EXPERIMENTAL AND COMPUTATIONAL METHODS FOR EXPLORING THE GLYCOME AND GLYCOPROTEOME

by

Shadi Toghi Eshghi

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

March, 2016

Abstract

Glycosylation – attachment of carbohydrates to proteins – is the most prevalent form of post-translation modification and responsible for protein structural and functional diversification. These carbohydrates, also known as glycans, are ubiquitous and play key roles in many biological functions such as signal transduction, protein folding and quality control, cell recognition and pathogen invasion. Aberrant glycosylation is associated with major human diseases like cancer, viral and bacterial infections and neurodegenerative diseases. The goal of this dissertation is to develop novel experimental and computational tools that could provide new insights into the pathological changes of glycosylation. A mass spectrometry based imaging technique is introduced for direct profiling of N-linked glycans from formalin-fixed paraffin-embedded tissue sections. Imaging of mouse brain coronal sections with this technique revealed significant differences between the glycomics profiles of the midbrain and brain cortex. Notably, fucosylation appeared to be more abundant in the cortex, while oligomannose structures and non-fucosylated glycans were more abundant in the midbrain. Moreover, mass spectrometry imaging of N-linked glycans was employed to differentiate glioblastoma tumor cells injected into a mouse brain from the surrounding normal tissue. To examine the protein hosts of glycans and the microheterogeneity of glycosylation, an algorithm and accompanying software tool were developed for site-specific identification of glycopeptides from mass spectrometry glycoproteomics data. Spectral library matching is introduced to assign the structures of intact glycopeptides based on their higher-energy collisional dissociation fragmented tandem mass spectra. Taking advantage of the power of spectral library matching, novel glycan modifications were exposed. Machine learning was applied to the spectral features

of glycopeptides to predict their glycosylation type. The application of the developed software tool, named GPQuest, was verified on recombinant glycoproteins and employed to study complex samples like prostate cancer cell lysates. GPQuest, powered with an easy to use graphical user interface, is made available online for the glycobiology community.

Advisors:

Hui Zhang, Ph.D., Associate Professor of Pathology, Johns Hopkins University

Xingde Li, Ph.D., Professor of Biomedical Engineering and Electrical and Computer Engineering, Johns Hopkins University

Thesis Committee (alphabetically):

David Goodlett, Ph.D., Professor and Isaac E. Emerson Chair of Pharmaceutical Sciences, University of Maryland School of Pharmacy

Kevin Yarema, Ph.D., Associate Professor of Biomedical Engineering, Johns Hopkins University

Note on Published Work

Content from the following chapters with minor changes has been peer-reviewed and published, and is included with permission:

Chapter 4. Imaging of N-linked glycans from formalin-fixed paraffin-embedded tissue sections using MALDI mass spectrometry (published in ACS Chemical Biology [1] and reprinted with permission. Copyright 2014 American Chemical Society)

Chapter 5. GPQuest: A spectral library matching algorithm for site-specific assignment of tandem mass spectra to intact N-glycopeptides (published in ACS Analytical Chemistry [2] and reprinted with permission. Copyright 2015 American Chemical Society)

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my advisor, Dr. Hui Zhang, for her patience, leadership, and unwavering support during the course of my PhD studies. She has been an exceptional mentor, who has taught me how to be a better scientist in the lab and a stronger person in life and for that I am forever grateful. I am also very thankful to my co-advisor, Dr. Xingde Li, for his scientific rigor and constant encouragement from the day I started my graduate studies. I wish to extend my appreciation to my thesis committee members, Dr. David Goodlett and Dr. Kevin Yarema for their insightful feedback and constructive suggestions over that past two years that has made this work much stronger.

I would like to thank the members of the Center for Biomarker Discovery and Translation for their continued support during the past five years. I am particularly grateful to the principal investigators, Dr. Daniel Chen, Dr. Zhen Zhang and Dr. Lori Sokoll for creating an outstanding research environment and to my colleagues, Punit Shah, Weiming Yang, Jing Chen, Lijun Chen, Shisheng Sun, Jered Passay, Stefani Thomas, Lily Chen, Jake Yang, and Yingwei Hu for so many collaborations, scientific discussions and joyful moments. I would also like to thank the members of GWEN, whose friendship has made this journey more pleasant.

I am forever indebted to my parents, Nasrin Nazari and Asghar Eshghi for their many sacrifices, to my brothers, Amin and Moien, who have been my champions ever since I can remember and to my husband, Iraj Hosseini, for his love and believing in me and for his many “inspirational speeches”. I am also grateful to the Gooyas for being my family away from home.

This dissertation and my graduate studies have been supported by the Siebel Scholarship, Programs of Excellence in Glycosciences, and grants from the National Institute of Health, National Cancer Institute and National Heart, Lung, and Blood Institute. I greatly appreciate these financial supports for making this research possible.

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

1.1 Overview

The conventional dogma of molecular biology, which explains the template-driven expression of proteins from DNA molecules, is central to our understanding of how biological systems develop and evolve [3]. Despite being the driving force behind the biological revolution of the mid 20th century, this dogma tends to overlook two major classes of biomolecules — lipids and carbohydrates — that play critical roles in a cell's normal development and survival [4] (Figure 1-1). In addition, one could argue that the relatively small number of genes and the great overlap between the genome of humans with those of other species and organisms implies that genes alone could not be responsible for the biological diversity that is observed in nature.

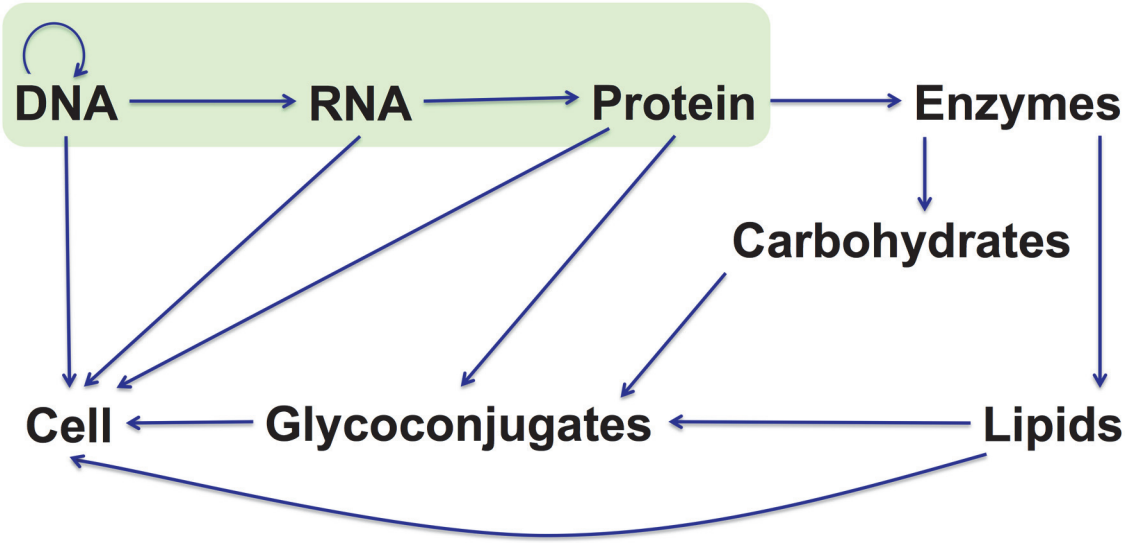


Figure 1-1. Central dogma of molecular biology.
The central dogma of molecular biology, highlighted in green, explains the flow of genetic information from DNA to RNA and subsequently proteins. This conventional view does not address the role of other major biomolecules, i.e. carbohydrates and lipids. In reality, many translated proteins form pathways that are responsible for biosynthesis of carbohydrates and lipids and assembling these biomolecules to build glycoconjugates, all of which play major roles in cell development and survival.

Protein post-translation modification (PTM) refers to covalent attachment of various molecules to proteins, mid- or post-biosynthesis, which affects both the structure and function of the host protein [5]. PTMs encompass a wide range of molecules with diverse chemical and physical properties and can be as small as a phosphate group in case of phosphorylated proteins or as large as glycosaminoglycans that are long linear polysaccharide chains on proteins forming proteoglycans [6]. Glycosylation is an enzyme-mediated PTM of carbohydrates on a modestly estimated half of the proteins and accounts for the highest level of protein diversification [4], [7]. Glycosylated proteins can be categorized into several classes depending on the structure of the carbohydrate and the site on which the attachment occurs. N-linked glycoproteins, O-linked glycoproteins, O-GlcNAcylated proteins and proteoglycans are the major forms of glycosylated proteins [4], [8].

Glycosylation plays a critical role in homeostasis of the cell and is involved in cell signaling, transcription regulation, protein folding and quality control, cell-cell and cell-extracellular matrix interactions, bacterial and viral infections, and cancer metastasis [4], [7], [9]. In fact, the outer surface of all living cells is covered with a layer of glycans forming the glycocalyx, which is the interface of the cell with the extracellular environment [4], [10]. Due to this, glycosylation is particularly crucial in multicellular systems, were the cell-cell interactions are essential to the development and survival of the organism [11]. In humans, aberrant glycosylation is linked to many diseases such as cancer [12]–[17], cardiovascular diseases [18]–[20], immune disorders [21], [22] and neurodegenerative diseases [23], [24]. During the course of these diseases, dysregulation of the glycan biosynthesis enzymes can lead to synthesis of abnormal glycan structures

that not only might reveal the underlying pathological events, but also could serve as potential drug targets for treating the disease or as biomarkers for early detection.

Studying glycosylation is one of the more challenging branches of molecular biology and chemistry owing to several factors. First, glycan structures are inherently complex. Unlike DNA, RNA and proteins, glycans are branched structures composed of several monosaccharide building units attached through glycosidic bonds. Each glycosidic bond has two possible stereoisomeric forms of α or β and the anomeric carbon of one monosaccharide can link to any of the hydroxyl groups on the second one [25]. The branching and diversity of linkages in glycosylation can result in complex structures, which pose a challenge to glycan analysis and synthesis. Furthermore, the glycan structure is not explicitly coded by the genome. The glycosylation machinery is utterly sophisticated, consisting of numerous glycosyltransferases and glycosidase enzymes that are encoded by more than 200 genes. The activities of these genes are affected by events that are specific to the cell type and development stage, not to mention the cell nutrition status. Glycosyltransferases and glycosidases construct the glycans by sequentially adding to or removing monosaccharides from the substrate. The non-template nature of glycosylation is an important reason why glycosylation studies have been historically impeded [4].

Glycans and glycoproteins are diverse molecules and subsequently diverse methods and tools are required to elucidate the structures of these biomolecules. The analytical methods for study of glycans and glycoproteins can be summarized as isolation and identification, sometimes followed by quantification. Isolation of glycans or glycoconjugates can be achieved through chemical immobilization followed by

enzymatic reactions [26] and chromatographic [27], [28] or affinity-based separation [29]. Glycan binding proteins, also known as lectins, are a prevalent, fast and effective means for epitope recognition and affinity-based extraction of glycoconjugates [30]. High-resolution structural analysis of glycans can be accomplished by NMR and mass spectrometry and these methods combined with derivatization with fluorophores or mass labels permit quantification of the glycans or glycoconjugates [4], [25], [31].

In recent years, there have been national and international efforts to promote the significance of glycosciences and the promising opportunities that it offers to research and patient care and to paint a roadmap as a guide for researchers in the field [9]. Scarcity of accessible analytical tools for glycans and glycoproteins has been identified as one of the obstacles for glycosciences to fully integrate with other branches of molecular biology [9]. The objective of this dissertation is to address some of the niches for analytical techniques in the field of glycosciences and provide novel tools to help researchers delve into glycomics and glycoproteomics. This was accomplished by developing experimental and computational tools for studying glycans and glycoproteins in simple and complex biological samples. These tools focus on mass spectrometry imaging techniques for direct profiling of glycans from tissue sections and bioinformatics methods for high-throughput site-specific analysis of glycoproteins. Applications of these tools are demonstrated in various biological samples, including recombinant proteins, cell lines, animal models and human tumor cells.

1.2 Organization of the thesis

The organization of this thesis is as follows: Chapter 2 provides an overview of protein glycosylation, and structural diversity, functional roles of glycosylation and significance

of glycosylation in human health. Chapter 3 presents the technologies and methods that are used for analysis of glycosylation with emphasis on mass spectrometry techniques. These two introductory chapters intend to provide the reader with sufficient background on glycosylation and present the concepts and nomenclature that is used in this dissertation.

Chapter 4 presents a mass-spectrometry based technique for imaging the distribution of N-linked glycans on formalin-fixed tissue sections. This technique permits high-throughput imaging of tens of glycans in one single experiment with superb specificity. Application of this imaging technique on mouse brain section revealed clear distinctions between the glycosylation patterns in midbrain versus the cortex. In addition, mass spectrometry imaging was used to differentiate tumor cells injected into healthy mouse brain in this study.

Chapter 5 and 6 focus on computational algorithms and software tools for high-throughput analysis of mass spectrometry glycoproteomics data. In chapter 5, spectral library matching is introduced for assignment of tandem mass spectra to intact N-glycopeptides structures and the application of this algorithm in analyzing cell lysate extracts of LNCaP prostate cancer cells is demonstrated. In chapter 6, the glycoproteomics algorithms are extended to include O-glycoproteins as well as N-glycoproteins. This was achieved by combining the spectral features of tandem mass spectra with machine learning to predict the glycosylation type for each tandem mass spectrum. Furthermore, GPQuest is introduced in this chapter, which is a software package developed in conjunction with this project for glycoproteomics analysis of

higher-energy collisional dissociation fragmented data of glycoprotein mixtures and biological samples.

Chapter 7 concludes this dissertation by providing preliminary data on the relevant ongoing studies and discussing future directions for this research.

Chapter 2. Background – Protein Glycosylation and Its Role in Health and Disease

Glycans are sugar chains and branched assemblies of monosaccharides that are post-translationally attached to more than half of the total human proteins and mediate their function. They coat the surface of all living cells, controlling their interactions with the extracellular environment through cell signaling and recognition. According to the central dogma of molecular biology, genes are first transcribed to RNA sequences, which in turn are translated to proteins. For proteins however, this is just the beginning. They further mature during protein folding, are modified by an array of post-translational modifications (PTM) and go through a quality control process. Glycosylation is the most prevalent form of PTM and affects the majority of the proteins. This chapter provides an overview of protein glycosylation, its functional roles and significance in human health and disease.

2.1 N-linked and O-linked protein glycosylation

The most common forms of protein glycosylation can be categorized into N-linked and O-linked classes. N-linked glycosylation occurs on NXS, NXT, and less commonly NXC and NXV sequences [32], where X can be any amino acid except proline. N-glycans have a common core of two N-acetylglucosamine (GlcNAc) and three mannose residues and can be subdivided into oligomannose, complex and hybrid structures as shown in Figure 2-1 [4]. N-glycosylation is particularly common on secreted and membrane-bound proteins on cell surface [33]. O-linked glycosylation occurs on S, T or less commonly Y residues on proteins [34]. O-glycan structures are more diverse than N-glycans and include subclasses of O-GalNAc, O-fucose and O-mannose, where for O-GalNAc – also known as mucin-type — glycans alone, 8 core structures have been observed [4]. Mucin-type glycoproteins, which are found in excess in mucous, contain S/T rich regions that

are heavily O-glycosylated. These regions usually are rich in proline as well, which subsequently facilitates the O-glycosylation of S/T residues [35]. Mucin-type glycoproteins are believed to provided immunity against pathogens by serving as decoys to their lectin receptors [36]. Both N- and O-linked glycosylation happen in Golgi and Endoplasmic Reticulum [8]. N-linked and O-linked glycans are usually adorned with fucose or sialic acid residues on the core GlcNAc or on the elongated branches, which facilitate the recognition of glycoproteins by lectins and antibodies [37]. Another form of protein glycosylation, called O-GlcNAcylation, is the attachment of a GlcNAc to the S or T residues on proteins by the O-GlcNAc transferase (OGT) enzyme and can happen in cytoplasm, nucleus or mitochondria. O-GlcNAcylation is a dynamic modification in nature and can be reversed by detachment of the GlcNAc from the peptide using the O-GlcNAcase (OGA) enzyme, which makes the S or T residue available for another potential round of modification by a phosphate group or GlcNAc [38]. Proteoglycans are another form of glycosylated proteins and constitute proteins linked to long chains of glycosaminoglycans via a xylose residue [6].

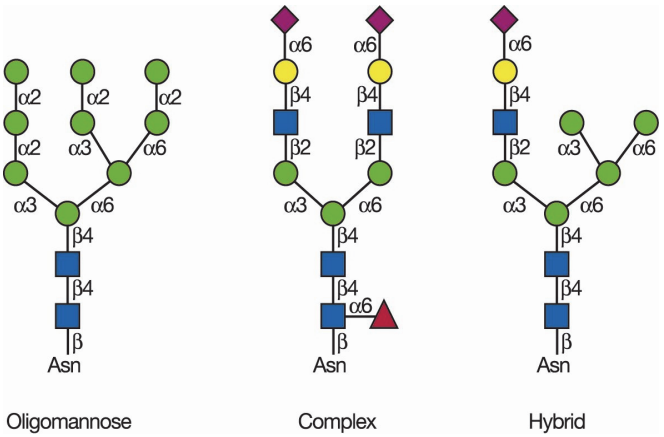


Figure 2-1. Types of N-glycans. N-linked glycans are connected to the asparagine residue on the protein backbone via an N-glycosidic bond. Typically, N-linked glycan structures can be sorted into oligomannose, complex or hybrid. (Reprinted from 'Essentials of Glycobiology', Cold Spring Harbor Laboratory Press [4] with permission. Copyright © 2009, The Consortium of Glycobiology Editors, La Jolla, California)

Dissimilar to proteins that fold inward, resulting in burying parts of the protein backbone, glycans tend to extend outward and occupy space. Considering the fact that glycans are large molecules oftentimes comparable in mass and size to their host peptide sequences, modification of proteins by glycans affects their 3-dimensional (3D) structure and consequently, their activity. They carry much of the information content of living systems, thus relating the genomics data with the observed phenotype [9]. A membrane-bound glycoprotein modified by N-linked and O-linked glycans is shown in Figure 2-2 [39]. To facilitate the illustration of glycans, each monosaccharide is represented by a shape and the key is provided in this figure. This thesis follows the standard presented in Figure 2-2.

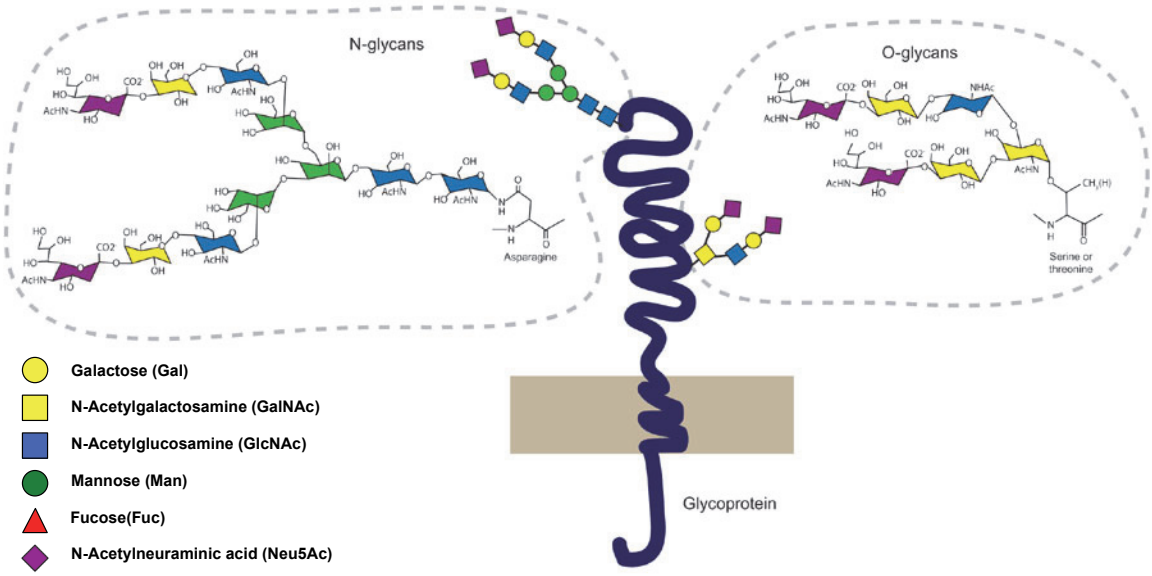


Figure 2-2. Schematic of a glycoprotein modified by N-linked and O-linked glycans. Monosaccharides are building blocks of glycans and depending on their chemical composition and 3D structure, are represented by different shapes. (Reprinted from 'Protein-glycan Interactions in the Control of Innate and Adaptive Immune Responses', Nature Immunology 9, 593- 601 (2008) [39] with permission. Copyright © 2008, Rights Managed by Nature Publishing Group)

2.2 Microheterogeneity of glycosylation

Unlike proteins, structures of glycans are not explicitly coded by the genome. In fact, protein glycosylation is determined by proteins involved in glycan biosynthesis pathways whose activities are affected by protein abundance and cell type-specific events. In addition, glycosylation at a specific glycosylation site is regulated by other factors such as substrate glycoprotein abundance, protein folding, cell type and its development and metabolic state. These factors result in what is called the microheterogeneity of glycosylation, where the occupancy of identical protein glycosylation sites by different glycan structures varies and each protein can be found in several diverse glycoforms [4], [40], [41]. Microheterogeneity plays key roles in the function of this post-translational modification [40], [42], [43] and therefore the ability to preserve this information is critical. For example, increased sialylation of glycans on IgG affects its anti-inflammatory properties [44]. Moreover, numerous studies have shown that during the progression of diseases, both glycans and glycoproteins can go through changes in their structures and abundance, suggesting that in fact changes in glycans or glycoproteins are not independent from each other [11], [40], [42], [45]. Therefore, site-specific glycoproteomics analysis could potentially lead to identification and characterization of a new class of biomarkers that are more specific to their underlying pathology, similar to fucosylated alpha-fetoprotein for detection of liver cancer [46]. For example, metastatic prostate cancer cells exhibit higher levels of fucosylation. However, this increased fucosylation occurs in a site-specific manner such that even glycosites on the same protein are affected differently in response to elevated levels of fucosyltransferase enzymes [47]. The ability to monitor the changes in the microheterogeneity of glycosylation induced by different pathological conditions is critical for diagnosis and

treatment. In addition, this topic is of particular interest in developing antibodies against glycoproteins and in the field of vaccine development [48].

2.3 Significance of glycosylation to human health

Glycans play vital roles in almost all biological processes (Figure 2-3 [7]), e.g. transcription regulation, cell proliferation, cell signaling, cell-cell and cell-extra cellular matrix interactions, bacterial and viral infection, protein degradation, inflammation and activation of the immune system [4], [7], [8]. Yet, the extent of damage imposed on the cell or organism as a result of specific alterations in glycosylation depends highly on factors like cell type, developmental stage and whether the experiment is *in vivo* or *in vitro*, and can vary from being undetectable to fatal [11]. For example, while knock down of GlcNAcT-1 glycosyltransferase gene is tolerated in cell lines, it exhibits severe phenotypes in mouse models [49].

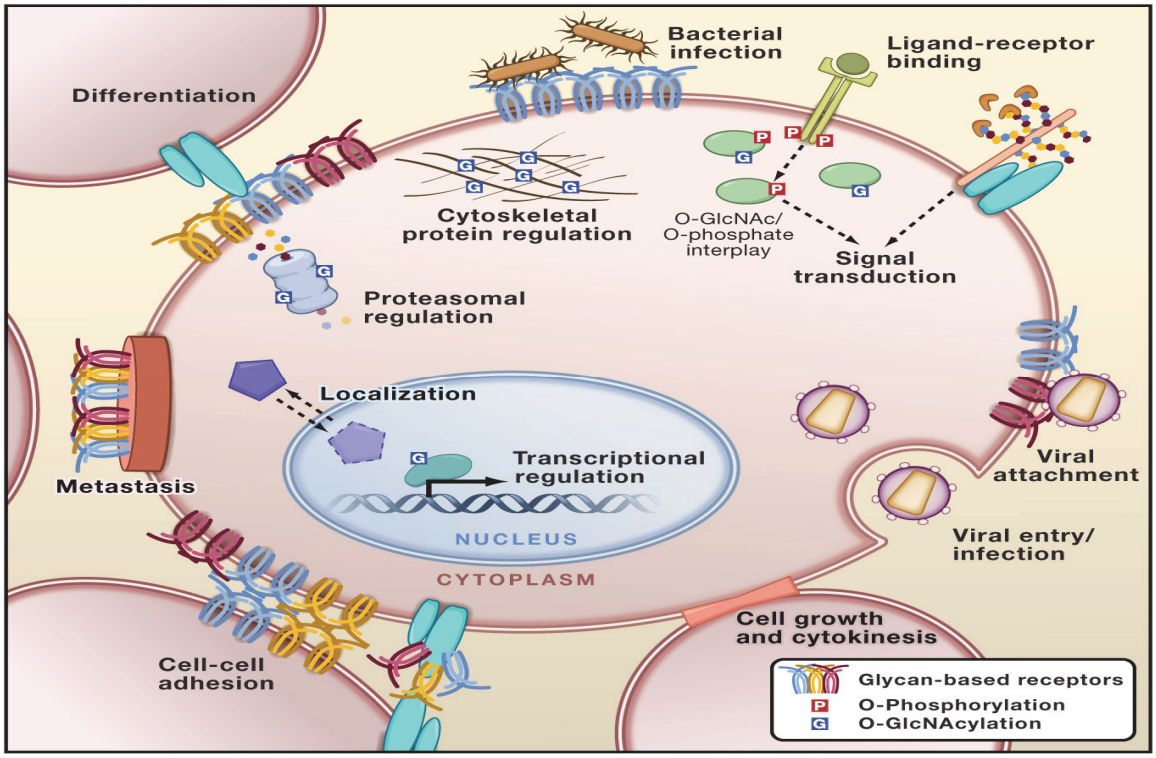


Figure 2-3. Functional roles of glycosylation.
Glycosylation in intertwined with almost all cellular functions from regulating the transcription to protein quality control, cell signaling and cell interactions with other cells or the extracellular environment. (Reprinted from 'Glycomics Hits the Big Time', Cell 143, 672–676 (2010) [7] with permission. Copyright © 2010 Elsevier Inc.)

Glycans are involved in almost all diseases that affects humans such as cancer, cardiovascular diseases, neurodegenerative diseases and disorders of immune system as observed and confirmed in numerous studies [4], [7], [11], [40], [50]–[52]. In cancer, abnormal glycosylation is a universal feature observed in many cancer cell types and tissues such as prostate, breast, ovarian, lung and liver cancers [12], [53]. Some of the most widely used clinical assays for cancer screening and diagnosis aim at detection of glycosylated proteins, among which are prostate-specific antigen and cancer antigen 125 [53]. The glycosylation profile of the cell undergoes changes during tumor formation and metastasis. For example, aberration in the cell surface glycans leads into more aggressive

cellular behavior and cancer progression and presence of truncated O-glycans on the cell surface is linked to aggressive cancer [15], [54]. Congenital disorders of glycosylation (CDG) are a class of inherited disorders caused by genetic mutations that result in defective glycosylation machinery or metabolism [55]. CDGs are very rare because they are oftentimes fatal in early development, and in case of survival, lead to severe phenotypes in childhood and early adulthood. In the immune system, lectins on the cell surface are capable of recognizing foreign glycans from bacteria and pathogens, thus protecting the cell from infection. Likewise in autoimmune disorders, deficiencies in the glycosylation biosynthesis pathway may construct abnormal and potentially immunogenic glycan epitopes [11] that provoke an immune response from the cell surface lectins. Glycans are also critical to endocytosis and cell trafficking. Therefore, deficiencies in glycosylation could disturb transportation of defective proteins and glycans to lysosome for degradation, resulting in cellular storage diseases [11].

Every year, more diseases are discovered to be attributed to deficiencies in glycosylation. Diagnosing diseases of glycosylation and understanding their biology has saved many patients from severe symptoms and improved their quality of lives. For example, a treatment as simple as taking oral mannose supplements alleviates the life-threatening symptoms of CDG-1b patients [56]. Expanding the toolbox of researchers in glycosciences is an essential step that could help reveal better diagnostics and therapeutic targets and therefore improve human health.

Chapter 3. Background – Overview of Analytical Methods for Studying Glycosylation

The past decade has witnessed dramatic progress in development of analytical tools and methods for studying glycosylation from biological sources. The glycosciences have greatly benefited from years of experience in protein analytics and many of the glycoproteomics analytical methods have originated from proteomics. Even though protein analytics have offered glycosciences a fundamental toolbox, numerous challenges remain unresolved. Since the complexity of glycans and glycoproteins surpasses that of proteins, additional analytical techniques tailored to these biomolecules are required. This chapter presents an overview of the existing technologies that are used for analysis of glycans and glycoproteins as well as the niches that offer opportunities for innovation and growth.

3.1 Experimental analytical tools for glycans and glycoproteins

Analysis of glycans and glycoconjugates starts with isolation of these analytes from the biological sources to improve the detection sensitivity and minimize interference from other analytes. Isolation and purification of glycoconjugates is followed by detection, identification and if applicable, quantification of these analytes [25], [31]. Lectins, solid phase extraction, endo- and exoglycosidases, nuclear magnetic resonance, chromatography, and mass spectrometry are among the most common techniques for isolation, detection and identification of glycans and glycoproteins.

3.1.1 Lectins

Lectins are naturally occurring glycan-binding proteins that recognize certain glycan epitopes with great specificity [30], [57]. Lectins are inexpensive and easy to use and therefore one of the most ubiquitous reagents for glycan analysis. They are commercially

available in different forms such as native, biotinylated for lectin histochemistry analysis, agarose-conjugated for immunoprecipitation and fluorophore-conjugated for microscopy and cytometry of glycans and glycoconjugates. In addition, lectin microarrays have been developed for quick and high-throughput glycan profiling and biomarker screening of multiple samples [58], [59]. Besides being inexpensive and easy to use, lectins are capable of differentiating glycan isomers and linkage types, which is a major hurdle in glycan structural analysis. For example, while it is not straightforward to distinguish different hexose sugars using mass spectrometry, concanavalin A (ConA) and ricinus communis agglutinin (RCA) bind mannose and galactose, respectively and can be used to tell them apart [30]. Lectins are versatile tools for glycan analysis, but they have certain limitations. Each lectin recognizes certain epitopes; consequently two different glycoforms sharing the same epitope cannot be distinguished using lectins alone. This property necessitates the application of additional orthogonal methods such as mass spectrometry, when detailed structure of the whole glycan or glycoconjugates is of significance.

3.1.2 Solid phase extraction of glycans and glycoproteins

Solid phase extraction has been applied for isolation of glycans and glycoconjugates to great effect [60]–[63]. In this method, glycoproteins are chemically immobilized on solid beads conjugated to functional groups e.g. carboxyl, aldehyde or amine groups via the glycan end, protein termini or protein side chains. The flexibility of the solid phase extraction technique allows for isolation of glycans or glycosite-containing peptides. For example, to analyze the glycan portion, Yang *et al* conjugated the N-termini of glycoproteins via reacting the protein amine groups with aldehydes on the bead surface

[26], [29]. On the contrary when the sequence of the glycosite-containing peptide is of interest, the glycoproteins can be conjugated on solid support on their glycan end by reacting the oxidized glycans to hydrazide functional groups on the beads, leaving the glycosite free for further enzymatic digestion [61], [63].

Solid support extraction retains the sample during clean up, thus minimizing sample loss and improving the sample preparation efficiency. Glycans are regularly derivatized with mass tags for quantification, or esterified or permethylated for sialic acid stabilization or linkage analysis [60], [64], [65]. Although derivatization is an integral part of glycomics analysis, it can result in contaminations and/or sample loss. Immobilization of glycoconjugates on solid phase allows for straightforward clean up after derivatization of the glycoconjugates and minimizes the sample loss due to washing steps, because the glycoproteins are attached to the solid support via a covalent bond. Following the sample clean up, glycans or glycosite-containing peptides of interest are released by enzymatic or chemical reactions, collected and purified for identification.

3.1.3 Endo- and Exoglycosidases

When the objective of a glycomics study is to analyze intact carbohydrates, the glycans can be released from glycoproteins using enzymatic or chemical reactions or a combination of both. Peptide-N-glycosidases are enzymes that cleave off the N-linked glycans from their host protein without breaking the glycosidic bonds between the monosaccharides. Due to lack of universal enzymes for release of O-linked glycans en bloc, alternative chemical reactions like beta-elimination or hydrazinolysis can be used instead [31]. Moreover, to attain detailed structural features of the glycans e.g. sugar isomers and linkage anomericity, single monosaccharides can be sequentially released

using exoglycosidase enzymes. Exoglycosidases remove the terminal monosaccharide from the glycans with superb specificity to not only the sugar type but also the linkage. Sequential treatment with an array of exoglycosidases is oftentimes combined with chromatography or electrophoresis for sequencing the glycan structure [66].

3.1.4 Nuclear magnetic resonance spectroscopy

Nuclear magnetic resonance (NMR) spectroscopy is a non-destructive technique for characterizing the structure of glycans and glycopeptides [67]. NMR spectroscopy provides extensive details about the structure of glycans. In particular, NMR is capable of revealing the anomericity of the examined monosaccharides [4] and is an effective tool for probing glycan-protein interactions. NMR spectroscopy requires milligrams of purified sample for operation, which is difficult to achieve, particularly in clinical samples. The low sensitivity of NMR is therefore the main hurdle in broad application of this technique. Nano-NMR probes are improved implements that could be used for analyzing few nano-moles of purified samples [67].

3.1.5 Chromatography

Owing to the complexity of glycan and glycoprotein samples, low abundance of glycoconjugates and diversity of glycan isomers, separation techniques are usually required to improve the detection sensitivity [68]. Liquid chromatography is routinely used for separation of glycans and glycopeptides.

Hydrophilic interaction chromatography (HILIC) takes advantage of the hydrophilic nature of glycans and glycopeptides to retain these analytes on the polar stationary phase. The glycans or glycopeptides are eluted from hydrophobic to hydrophilic using high to

low gradients of acetonitrile solutions [27], [28], [43], [69]. Porous graphitized carbon (PGC) is another type of support commonly used for purification and separation of glycans and glycopeptides [28], [29], [70]. PGC offers broad working pH range, stability at high temperatures and is chemically non-reactive [68]. Consequently, it is a practical tool in harsh experimental conditions. Reverse-phase chromatography is typically used for separation of glycans that have become more hydrophobic due to labeling with fluorescent tags [31], [71]. Glycans do not absorb ultraviolet light; therefore labeling glycans with fluorescent tags improves their limit of detection. Labeling kits for these tags, for example 2-aminobenzoic acid, 2-aminobenzamide and 3-(acetylamino)-6-aminoacridine, are commercially available.

Chromatographic separation usually precedes mass spectrometry analysis of glycans and glycoproteins [71]. However, it could also serve as a means to identify the glycans and characterize their structure. The retention time of glycans on a given column can be estimated as a function of their structural properties and number of monosaccharides [72]. Alternatively, comparing the results with a library of chromatograms of known glycan standards would help in characterizing the glycan.

3.1.6 Mass spectrometry

Mass spectrometry is a powerful and sensitive high-throughput tool for analysis of glycans and glycoproteins, achieving a limit of detection down to the attomole range. High resolution mass analyzers such as time-of-flight (TOF) and orbitrap in combination with tandem mass spectral analysis provide superb specificity. Mass spectrometry approaches commonly follow the liquid chromatography separation of glycoconjugates

samples and can be applied for quick profiling as well as detailed structure determination of glycans and glycoproteins.

Matrix-assisted laser desorption ionization mass spectrometry (MALDI-MS) is one of the most common ionization techniques in glycomics applications [73], [74]. In MALDI mass spectrometry, glycans or glycoproteins are covered with a matrix such as α-cyano-4-hydroxycinnamic acid (CHCA) or 2,5-dihydroxybenzoic acid (DHB) and air-dried. The matrix absorbs the laser energy and transfers it to the analytes, thus creating a soft ionization effect, which protects the glycans or glycopeptides from unwanted fragmentation. MALDI creates predominantly singly charged ions and therefore the mass spectra generated by this technique are relatively straightforward to analyze. MALDI ionization often results in loss of sialic acid and therefore sialylated glycans are often permethylated or derivatized before MALDI-MS [65], [75]. MALDI ionization is routinely coupled to time-of-flight mass analyzers. MALDI-TOF/TOF instruments are particularly successful for analysis of high molecular weight biomolecules including glycans and glycoproteins [76]. MALDI-MS has been effectively applied for glycomics analysis in numerous studies [76]–[81].

Electrospray ionization (ESI) is another technique commonly used for analysis of glycans and particularly glycopeptides [74]. In ESI, high voltage is applied to the solution creating a fine spray of charged droplets, which turn into gaseous phase that contain the ions [82]. ESI creates ions bearing multiple charges, leading to more complicated mass spectra. So, analyzing the mass spectra generated by ESI usually requires more effort. Nano-ESI with high mass resolution analyzers such as orbitrap is particularly common for analysis of glycopeptides [43], [83]–[85]. Ion traps are also integrated in these mass

spectrometers for improving MS^n analysis by trapping enough ions for high quality fragmentation [76].

Tandem mass spectrometry analysis is key in identification of glycans and glycopeptides using mass spectrometry. Depending on its mechanism of action, fragmentation of glycans and glycopeptides reveals unique details about the structure of precursor ions. Several fragmentation techniques have been utilized in glycomics studies, such as electron transfer dissociation (ETD) [74], [86], [87], collision induced dissociation (CID) [78], [88], [89] and higher-energy collisional dissociation (HCD) [43], [47], [90], [91]. These methods often provide complementary information about the structure of glycans and glycopeptides. For example, ETD results in breaking the peptide bonds in the glycopeptide structures leaving the post-translation modifications intact. Therefore, it is an effective tool for creating the fingerprint of the peptides and determination of the modification sites. In contrast, CID breaks glycosidic bonds, generating glycan fingerprints and therefore is ideal for detailed structural analysis of the glycan portion of the glycopeptide. HCD results in partial fragmentation of both the peptide and glycosidic bonds. Typically, multiple fragmentation methods work in tandem for complete structural determination of glycopeptides [86], [89], [92].

3.2 Mass spectrometry based glycoproteomics

High-throughput investigation of glycosylation is oftentimes accomplished through mass spectrometry analysis. Until a few years ago, these studies mainly involved enrichment or isolation of glycopeptides using hydrazide chemistry, lectin affinity techniques or hydrophilic interaction chromatography, followed by detaching the glycan portion from the protein using chemical or enzymatic methods. The glycans and glycosite-containing

peptides were then isolated and examined separately [63], [70], [78], [93], [94], which led to loss of the microheterogeneity information of glycosylation.

In the past few years, advances in mass spectrometry techniques and instrumentation have made it possible to study and identify intact glycopeptides from recombinant proteins, glycoprotein cocktails or complex biological samples such as cells, serum and tissues [2], [43], [91], [95]–[99]. In shotgun glycoproteomics, the glycoprotein extract is digested using protease enzymes, most commonly trypsin. The resulting mixture of peptides and glycopeptides is separated using high-performance liquid chromatography (HPLC) and the eluted analytes are ionized and analyzed by mass spectrometry. For tandem MS analysis, certain mass spectral peaks are purified in the gas phase for fragmentation to generate fragment ions from specific precursor ions. The subsequent tandem mass spectra represent glycan specific fingerprints. While traditional ion trap mass analyzers have suffered from a low mass cutoff that obscured the low m/z region, newer Orbitrap analyzers have overcome this limitation. This is of particular importance in analyzing HCD fragmented glycopeptides, where detection of low mass monosaccharide ions is crucial for identification of glycopeptides. The significance of these advances is that for the first time, we are able to preserve and study the microheterogeneity of glycosylation in a high-throughput manner, meaning that not only can we characterize and quantify the glycosite-containing peptides and glycans, but we can also scrutinize glycosylation in a site-specific manner. On one hand, the advancements in technologies have enabled us to study the microheterogeneity of glycosylation in a high-throughput fashion. On the other hand, these instruments are capable of generating data at rates as high as 1 gigabyte/hour resulting in massive

amounts of glycoproteomics data being accumulated. Development of software tools and algorithms for high-throughput interpretation of glycoproteomics data is an open area of research. Database search and precursor mass matching, spectral library matching, and less commonly de novo sequencing can be used to assign the tandem mass spectra of fragmented glycopeptides to their corresponding structures [100]. Different fragmentation mechanisms generate distinct glycopeptide tandem spectra and therefore require specifically designed algorithms. For example, CID breaks the glycosidic bonds of the glycan portion, whereas ETD breaks the peptide bonds instead, resulting in very different tandem mass spectra representing the same glycopeptide structure. These patterns must be taken into account when designing or choosing algorithms for analysis of glycoproteomics data. Glycoinformatics has made great strides in the past decade and the glycobiology field has experienced an outburst in the number of software tools for automated identification and analysis of glycan and glycoconjugates from different data sources. Integration of bioinformatics tools into glycomics and glycoproteomics platforms has been instrumental in taking advantage of the advances in biotechnology and instrumentation. The most recent software tools for glycoproteomics are thoroughly reviewed by Hu *et al* [100].

Chapter 4. Imaging of N-Linked Glycans from Formalin-Fixed Paraffin-Embedded Tissue Sections Using MALDI Mass Spectrometry

4.1 Summary

Aberrant glycosylation is associated with most of the diseases. Direct imaging and profiling of N-glycans on tissue sections can reveal tissue-specific and/or disease-associated N-glycans, which not only could serve as molecular signatures for diagnosis but also shed light on the functional roles of these biomolecules. Mass spectrometry imaging (MSI) is a powerful tool that has been used to correlate peptides, proteins, lipids, and metabolites with their underlying histopathology in tissue sections. Here, we report an MSI technique for direct analysis of N-glycans from formalin-fixed paraffin-embedded (FFPE) tissues. This technique consists of sectioning FFPE tissues, deparaffinization, and rehydration of the sections, denaturing tissue proteins, releasing N-linked glycans from proteins by printing peptide-N-glycosidase F over the sections, spray-coating the tissue with matrix, and analyzing N-glycans by matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS). Brain sections from a C57BL/6 mouse were imaged using this technique at a resolution of 100 μm. Forty-two N-glycans were analyzed from the mouse brain section. The mass spectrometry images were used to study the relative abundance of oligomannose, nonfucosylated, and fucosylated complex N-glycans in different brain areas including isocortex, hippocampal formation, and brainstem and specific glycans associated with different areas of the brain were identified. Furthermore, glioblastoma tumor xenografts in a NOD/SCID mouse were imaged. Several glycans with differential expression in tumor versus normal brain tissues were identified. The MSI technique allows for imaging of N-glycans directly from FFPE sections. This method can potentially identify tissue-specific and/or disease-associated

glycans coexpressed with other molecular signatures or within certain histological structures.

4.2 Introduction

Imaging of tissue N-glycans is an essential, yet less-explored tool for studying their functions. In contrast to conventional glycan profiling assays, where the tissue is first homogenized, imaging focuses on studying the glycosylation heterogeneity in pathologically or structurally different regions of the tissue. Therefore, it can provide an invaluable means to understanding the roles of N-glycans in the physiology and molecular pathology of the diseases.

Histochemical staining using lectin is by far the most common method for visualization of glycosylation from formalin-fixed paraffin-embedded (FFPE) tissue sections. For instance, Concanavalin A (ConA) and Aleuria Aurantia Lectin (AAL) are two of the lectins that are used in histostaining of oligomannose and fucosylated glycans, respectively [30]. Despite its contributions to the studies on glycosylation in pathological tissue sections, this technique is limited in many ways. First, lectins provide minimal structural information about the stained epitopes. For example, ConA can bind internal and non-reducing terminal α-mannose [101], thus staining a variety of oligomannose structures. Therefore, the ConA staining cannot specify the structure of the glycans. In addition, they often fail to differentiate between different glycan subgroups. For instance, AAL can bind the fucose residues on both N-linked and O-linked glycans. Second, due to steric hindrance of different glycan epitopes, lectin histostaining is limited to very few lectins at a time on each tissue section, thus making multiplex glycan imaging very challenging. Third, histostaining methods lack quantification accuracy. Compared to the

lectins, glycan antibodies are more specific with respect to the glycan determinants. However, the number of currently developed monoclonal antibodies for glycans is far from covering the width of the mammalian glycans [9]. Therefore, alternative imaging techniques are essential to complement the information acquired from histostaining assays.

Mass spectrometry imaging (MSI) has been previously applied for spatially-resolved profiling of proteins, lipids, small molecule metabolites and drugs from tissue sections [102]–[106]. Owing to the high sensitivity and specificity of mass spectrometric analysis, MSI has overcome some of the challenges of conventional histostaining techniques. Unlike affinity-based detection methods such as immunohistochemistry staining, where the detection relies on some understanding of the analyte of interest, MSI does not require any *a priori* knowledge of the glycans. This attribute, which is a unique characteristic of MSI, is particularly desirable for discovery research. In addition, hundreds of analytes can be detected and identified from one single mass spectrometry experiment, resulting in high-content molecular profiling with spatial information in tissue sections. Furthermore, MSI can be combined with quantitative and semi-quantitative mass spectrometry analysis techniques to facilitate quantitative imaging of different analytes directly from tissue sections. The recent advances in mass spectrometry techniques and data interpretation have significantly pushed the limits of glycomics studies [107]. However, the low ionization efficiency of native glycans compared to other macromolecules such as proteins and lipids makes them challenging to study in complex mixtures. Therefore, glycans are often isolated from extracts of biological samples and chemically modified, e.g. permethylated, before mass spectrometry analysis [108], [109].

In this glycomics procedure, the spatial information of glycans is lost due to homogenization of the sample. Recently, we demonstrated that glycans released from glycoproteins that were immobilized on a solid phase could be directly analyzed by mass spectrometry. This method does not require further purification of glycans to remove interferences from proteins and peptides [26], [29]. This attribute is crucial for development of a glycomics study platform for imaging of glycans from samples with the high biological complexity of the tissues.

In this study, we developed a mass spectrometry-based platform for imaging of N-linked glycans from formalin-fixed paraffin-embedded (FFPE) tissue sections. In this technique, FFPE tissue sections are mounted on glass slides, which results in immobilization of the tissue glycoproteins on the slides. N-linked glycans are then selectively released from glycoproteins of tissue section by applying PNGase F enzyme using a microarray printer. The matrix-coated slides are subsequently analyzed by MALDI-MS. The acquired mass spectral images show the distribution of the N-linked glycans over the tissue section. Imaging of the N-linked glycans from FFPE coronal mouse brain tissue sections using this method revealed the spatial distribution of 42 N-linked glycans. In addition, the results showed that N-glycans are present in all regions of the brain. However, certain modifications are more abundant in particular brain structures. For example, brainstem (BS) is richer in oligomannose and non-fucosylated complex N-glycans, while the majority of the fucosylated N-glycans are more abundant in isocortex (IsoCTX) and the hippocampal formation (HPF). These observations were also compared with histostaining of adjacent tissue sections with AAL and ConA lectins. In addition, glioblastoma brain tumor xenografts from a NOD/SCID mouse were imaged by MALDI-

MS. Based on the acquired ion images, several N-glycans with differential expression in tumor versus adjacent normal tissues were distinguished, most of which were more abundant in the tumor.

The N-linked glycan mass spectrometry imaging platform not only helps identify the N-linked glycans directly from FFPE tissue sections, but also determines the spatial distribution of unique glycan structures over the tissue. This technique provides complimentary information to the traditional histostaining methods, which is essential to fully characterize the functional and pathological roles of N-linked glycosylation in tissues.

4.3 Methods

4.3.1 Materials and reagents

Antigen retrieval buffer was purchased from R&D Systems (Minneapolis, MN). Peptide-N-Glycosidase F (PNGase F) was from New England Biolabs (Ipswich, MA), Dithiothreitol (DTT), Maltoheptaose (DP7), 2,5-dihydroxybenzoic acid (DHB) and N,N-Dimethylaniline (DMA) were purchased from Sigma Aldrich (St Louis, MO). Biotinylated AAL and ConA lectins and the ABC-Elite kit were from Vector Labs (Burlingame, CA). Peroxidase blocking reagent was from Dako (Glostrup, Denmark).

4.3.2 Animal methods

A male C57BL/6 mouse from Jackson Laboratory (Bar Harbor, ME) was used for this study. It was housed in an animal facility with access to water, food and libitum. The mouse was euthanized at 20 weeks of age by harvesting organs and tissues: heart, aorta, kidney, liver, brain and spleen, under anesthesia with Ketamine/Xylazine (100mg/10mg

per kg IP). For imaging of the mouse brain tumor using MALDI mass spectrometry, 10^6 primary human glioblastoma cells (NS276) were stereotactically injected into the right striatum of an 8-week old, NOD/SCID male mouse (Charles River Laboratories) as previously described [110]. Four weeks following the tumor implantation, the mouse was sacrificed and the brain was extracted. These experiments were approved by the Johns Hopkins University Institutional Animal Care and Use Committee (Protocol numbers MO11M492 and MO12M195).

4.3.3 Mouse brain tissue fixation and embedding

The mouse brain tissues were fixed in 10% formalin for 48 hours after dissection. Following dehydration, the fixed brain tissues were embedded in paraffin. The samples were sectioned at thickness of 5 μm. The normal brain tissue sections were mounted on indium tin oxide (ITO) coated glass slides (Delta's Technologies, Loveland, CO), while the tumor brain sections were mounted on positively charged glass slides. The slides were stored at room temperature for a maximum of one month until use.

4.3.4 Tissue preparation

The FFPE tissue sections were deparaffinized by 3 xylene washes, 15 minutes each. Subsequently, they were rehydrated in graded ethanol solutions of 100, 95, 70 and 50%. To denature proteins on tissue slides, antigen retrieval procedure was performed by baking the tissue sections in the basic antigen retrieval buffer, pH 9.0 (R&D Systems) for 20 minutes. This was followed by protein denaturing in 40 mM DTT buffer and steaming for 10 minutes. Pretreatment of the tissue with denaturing reagents improves the deglycosylation significantly. The tissue was then briefly washed with 1% formic acid, 1

M sodium chloride and deionized water. The tissue was further washed with 15 mM ammonium bicarbonate buffer (pH ~8.0) for 20 minutes to equilibrate the pH of the section before applying the PNGase F enzyme. One additional section was prepared using this method to serve as the no enzyme control.

4.3.5 Deglycosylation and matrix deposition

PNGase F is the enzyme that cleaves the N-linked glycans from the attached proteins and peptides. 1 M ammonium bicarbonate buffer was added to PNGase F to bring the final pH to 8, close to the optimal pH for deglycosylation by PNGase F. 1mM DP7 was spiked into the enzyme solution as an internal standard to obtain a final concentration of 70 μM. PNGase F mixture was printed over the tissue section using an automated microarrayer (BioRobotics MicroGrid, Isogen Life Science, De Meern, The Netherlands) at 200 μm spacing. The robotic application of the enzyme using the microarray printer not only creates a uniform array of localized enzyme over the tissue, but also requires significantly less amount of PNGase F. The PNGase F-printed tissue section was incubated in a humidity chamber (maximum humidity of 80%) at 37°C overnight. The matrix was prepared by dissolving 120 mg of DHB into 1 mL of 50% acetonitrile, 0.1 mM sodium chloride followed by addition of 20 μL of DMA. The matrix solution was uniformly sprayed over the sample using an artistic airbrush (Aztek 470, Testors, Rockford, IL) according to the method previously described [111].

4.3.6 MALDI imaging

After the sample was air-dried, it was analyzed by a MALDI-QIT-TOF mass spectrometer (AXIMA Resonance, Shimadzu, Kyoto, Japan). The Launchpad software

was used to specify the mass analysis parameters such as the mass range, the laser intensity, scanning area and the spatial resolution. The mass spectral images were acquired in positive mode with 20 shots per profile at laser intensity of 130 in the mass range greater than 1170 Da, by scanning the laser at a spatial resolution of 100 μm. The image resolution was constrained by the acquisition time. At this resolution and with the specified settings, it took about 11 hours to image every cm^2 area of the section. The raw data files were converted to imaging files (.img) using the Launchpad software and the images were visualized and analyzed in MATLAB (Mathworks, Natick, MA).

4.3.7 Lectin histostaining

Followed by deparaffinization in 3 xylene washes and rehydration in graded ethanol, the endogenous peroxidase activity was blocked. The lectin (AAL or ConA) was diluted to a final concentration of 20 μg/milt in Dulbecco's phosphate buffered saline (DPBS) and incubated with the tissue for 30 minutes at RT. ABC-Elite kit was used for detection of the biotinylated lectin according to the instructions. Diaminobenzidine was applied as the chromogen to visualize the tagged glycans. The tissue was then counterstained with haematoxylin and cover slipped.

4.4 Results and discussion

Mass spectrometry imaging of glycans relies on enzymatic release of N-glycans from the proteins that had been immobilized on the glass slide. The attachment of proteins to solid phase minimizes their interference with the glycan mass spectral signal. The glycan imaging consists of multiple steps, including deparaffinizing and rehydrating the FFPE section followed by antigen retrieval to recover the protein reactivity thus improving the

efficiency of the PNGase F digestion. After equilibrating the pH of the tissue section followed by air-drying, PNGase F is printed over the section in a grid at a spatial resolution of 100 µm using a microarrayer. The PNGase F-printed tissue is then incubated in a temperature-controlled humidity chamber to complete deglycosylation. The matrix solution is sprayed over it and the matrix-coated section is imaged by MALDI-MS. Figure 4-1 shows a representative schematic of the workflow for imaging of N-linked glycans from FFPE tissue sections.

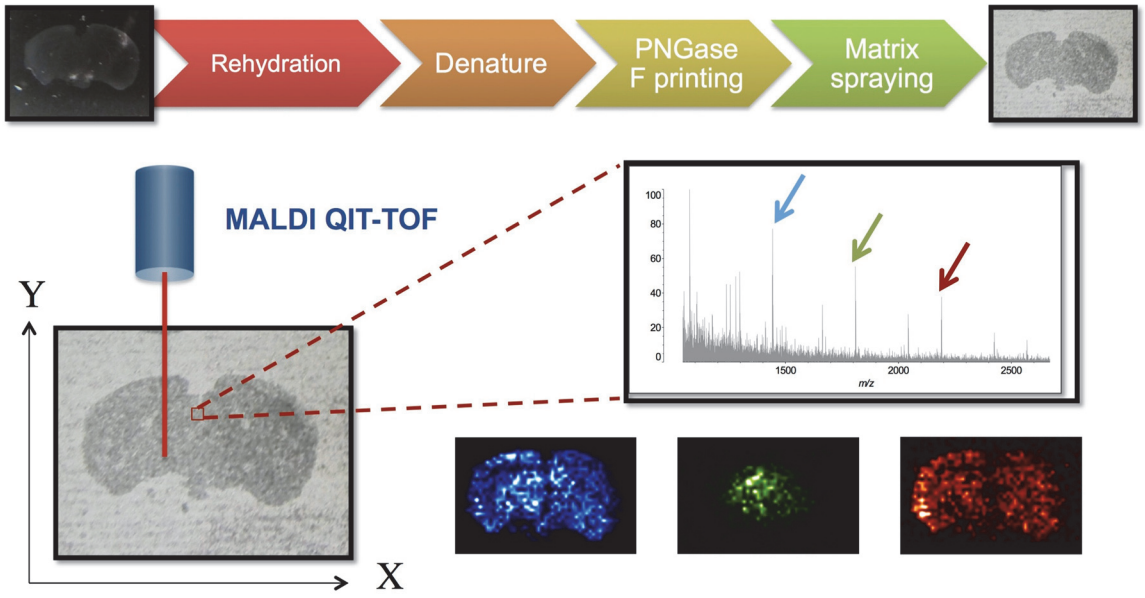


Figure 4-1. Schematic workflow of mass spectrometry imaging of N-linked glycans from FFPE sections.

An FFPE tissue section is first deparaffinized and rehydrated. The tissue proteins are denatured by treatment with a basic antigen retrieval buffer and baking in a water bath for 20 min followed by incubation in a 40 mM DTT solution. To preserve the spatial information on the glycans, a microarray printer is used to apply the PNGase F on the tissue in a grid. The section is incubated in a humidity chamber at 37 °C overnight. After air-drying the tissue, DHB matrix is sprayed over the section using an artistic airbrush followed by analyzing by MALDI-MS. The major difference between a conventional MALDI analysis and an imaging experiment is that here, the tissue is raster scanned by the laser in the x and y directions and mass spectra are acquired for each pixel on the tissue. At this point, by mapping the intensity of various peaks as a function of location, ion images can be generated for each glycan structure detected in the mass spectra. The

ion image corresponding to each mass spectral peak from the MALDI-MS spectra is shown in a different color.

4.4.1 Direct analysis of PNGase F-released N-glycans from FFPE tissue section using MALDI-MS

To determine whether glycans could be released and directly analyzed by MALDI-MS from glycoproteins immobilized on conductive slides, mouse brain coronal sections were analyzed. PNGase F was printed over the right half of a section, while buffer was printed over the other half at spacing of 100 µm as a negative control. One mm^2 area of each of the PNGase F negative and positive parts of the brain was imaged using MALDI-MS separately. Figure 4-2 shows the mass spectra corresponding to the PNGase F-negative (Figure 4-2A) and PNGase F-positive (Figure 4-2B) parts of the brain in the mass range of 1450 – 2400 Da. The mass spectral peaks in the bottom panel represent the N-glycans that are released from the glycoproteins in the PNGase F-treated section.

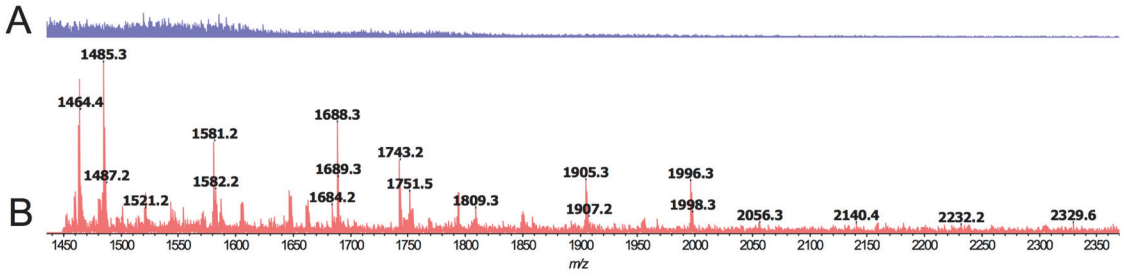


Figure 4-2. Direct analysis of N-glycans released by PNGase F from FFPE tissue section using MALDI-MS.
(A) PNGase F negative and (B) PNGase F positive. PNGase F is printed over the mouse brain coronal section at 100 µm spacing. The enzyme is printed on one-half of the tissue section, while the other half is treated with only the buffer. Mass spectra are acquired for a 1 mm^2 area on each brain half using MALDI-MS. The mass spectrum of the PNGase F negative part shows very low signal, while several N-glycan peaks are detected on the PNGase F positive sample.

To further ensure that the mass spectral signal corresponds to the glycans, collision-induced dissociation (CID) tandem mass spectrometry (MS/MS) was performed on the mass peaks in the MS1 spectrum. While not all of the mass spectral peaks were high enough to generate high-quality tandem MS spectra, the majority possessed the glycan signature mass differences of 162.05 and 203.08 corresponding to masses of a hexose (Hex) and N-acetylhexosamine (HexNAc) from N-glycans, respectively. The tandem MS spectra of two selected peaks corresponding to H7N2F0 and H4N4F1 are depicted in Figure 4-3 and Appendix Table 4-3. For comparison, glycomics analysis was conducted on mouse brain tissue extracts using glycoprotein immobilization for glycan extraction, which is described in detail previously [26]. The tandem MS spectra for the low-intensity peaks in the MALDI imaging spectra were acquired from the mouse brain extracted glycans to confirm their composition (Table 4-3). The fragment ions were manually assigned to the mass spectra with the help of the GlycoWorkbench fragmentation tool [112].

4.4.2 Identification of N-glycans from C57BL/6 mouse brain sections

Analyzing the tissue section with MALDI-QIT-MS resulted in detection and identification of 42 N-linked glycans (Table 4-1), where 30 (71.4%) of them were fucosylated and 7 (16.7%) of them were non-fucosylated complex glycans. All of the five oligomannose glycans (Man5, Man6, Man7, Man8, and Man9), constituting the 11.9% of the detected glycans, were identified. A list of the detected glycans is given in Table 4-1, where each N-glycan is depicted by its number of hexose (H), N-acetylhexosamine (N) and fucose (F) residues. The glycan composition identification was performed by first matching the peak mass with a database of all possible mammalian N-glycan compositions, and then refined by comparing the results with the Consortium for Functional Glycomics databases and the literature to remove the biologically irrelevance matches. Last, the glycan composition assignment was confirmed by evaluating the corresponding tandem MS spectra. Fucosylation could happen at the core or at the non-reducing ends of the glycans. Heavier, more branched glycans were generally lower in abundance and

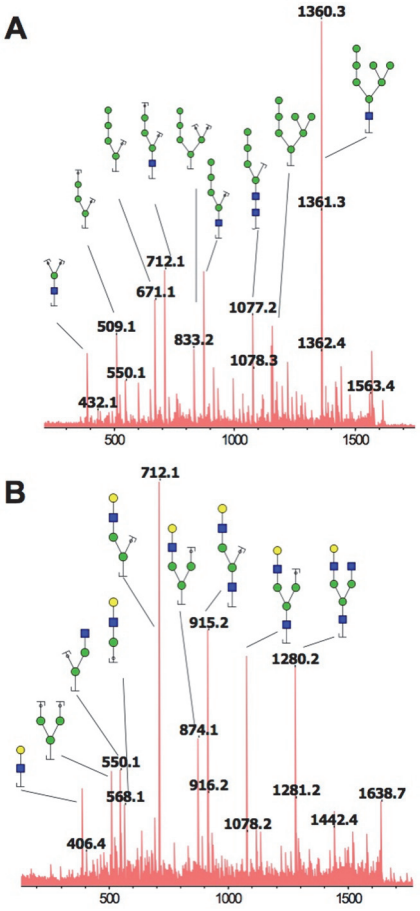


Figure 4-3. Examples of CID tandem MS spectra of the detected glycan peaks.

(A) H7N2F0 and (B) H4N4F1. Mass differences of 162.05 and 203.08 in tandem MS spectra are characteristic of glycans, which occur due to loss of Hex and HexNAc. The fragment ions are assigned to the mass spectra using the GlycoWorkbench tool. The acquired tandem MS spectra further support the specificity of this imaging technique to N- glycans.

harder to detect. The largest identified N-linked glycan was a highly branched structure with $[M+Na]^+$ theoretical mass of 2669.967 Da. In this experiment, however, sialylated glycans were missing from the spectrum. One possible explanation for this observation is the loss of sialic acid residues during mass spectrometry analysis by MALDI [81], [113]. The identified glycans were compared with mouse brain glycans reported on the Consortium for Functional Glycomics (CFG) database (http://www.functionalglycomics.org). The CFG database contains 31 unique non-sialylated N-glycan peaks in wild type C57BL/6 mouse brain tissue, 27 (87.1%) of which were detected in this experiment. These 27 glycan peaks are marked with a tick on the "Previously reported" column in Table 4-1. From the 15 N-glycan peaks that were identified in this study but not listed on the CFG data, 5 N-glycan peaks were identical to the de-sialylated counterparts of sialylated N-glycans in CFG mouse brain database. These 5 N-glycan peaks are marked with a cross on the "Previously reported" column in Table 4-1. The remaining 10 N-glycan masses had been previously reported in human plasma and serum samples [114], [115], 4 of which had also been detected in mouse brain tissue in a study conducted by Hu *et al* [116] or in whole rat brain tissue in another study conducted by Chen *et al* [117]. These 10 glycan peaks are marked by a plus sign on the "Previously reported" column in Table 4-1. This observation further supports that N-glycans can be directly released and identified from glycoproteins immobilized on slides. However, unprotected sialic acid residues have been lost during the acquisition of the mass spectra using MALDI [81], [113].

Table 4-1. The detected N-glycans from mass spectrometry imaging of mouse brain sections. Each glycan in the table is shown by its number of the hexose (Hex or H), N-acetylhexosamine (HexNAc or N) and fucose (Fuc or F) residues. Among the 42 detected N-glycans, 27 had been reported in the CFG database of wild type mouse brain previously

(http://www.functionalglycomics.org). Of the remaining 15 N-glycans, 5 most likely belong to sialylated glycans reported in the CFG mouse brain database that lost their sialic acid residues during sample preparation or ionization of the analytes. These 5 glycans are marked by a cross on the last column. With the exception of these 5 glycans, all of the remaining 10 detected N-glycans had been reported in human plasma and serum samples in earlier studies [114], [115]. Four of these 10 remaining glycans had also been reported in other mouse and rat brain tissue [116], [117].

#	Symbol	Hex (H)	HexNAc (N)	Fuc (F)	Theoretical mass $[M+Na]^+$	Detected mass $[M+Na]^+$	Previously reported	*In situ* MS/MS identification
1	H5N2F0	5	2	0	1257.4231	1257.343	√	Yes
2	H3N3F1	3	3	1	1282.4548	1282.368	+	Yes
3	H4N3F0	4	3	0	1298.4497	1298.387	+	Yes
4	H3N4F0	3	4	0	1339.4763	1339.417	+	Yes
5	H5N2F1	5	2	1	1403.4810	1403.429	+	Yes
6	H6N2F0	6	2	0	1419.4759	1419.405	√	Yes
7	H4N3F1	4	3	1	1444.5076	1444.413	+	Yes
8	H5N3F0	5	3	0	1460.5025	1460.401	+	Yes
9	H3N4F1	3	4	1	1485.5342	1485.456	√	Yes
10	H4N4F0	4	4	0	1501.5291	1501.446	+	Yes
11	H3N5F0	3	5	0	1542.5557	1542.473	+	Yes
12	H7N2F0	7	2	0	1581.5287	1581.446	√	Yes
13	H5N3F1	5	3	1	1606.5604	1606.455	√	Yes
14	H4N4F1	4	4	1	1647.5870	1647.482	√	Yes
15	H5N4F0	5	4	0	1663.5819	1663.476	√	Yes
16	H3N5F1	3	5	1	1688.6136	1688.508	√	Yes
17	H8N2F0	8	2	0	1743.5815	1743.498	√	Yes
18	H6N3F1	6	3	1	1768.6132	1768.494	√	No
19	H4N4F2	4	4	2	1793.6449	1793.523	√	Yes
20	H5N4F1	5	4	1	1809.6398	1809.524	√	Yes
21	H4N5F1	4	5	1	1850.6664	1850.579	√	Yes
22	H9N2F0	9	2	0	1905.6343	1905.517	√	Yes
23	H5N4F2	5	4	2	1955.6977	1955.562	√	Yes
24	H6N4F1	6	4	1	1971.6926	1971.554	√	No
25	H4N5F2	4	5	2	1996.7243	1996.588	√	Yes
26	H5N5F1	5	5	1	2012.7192	2012.541	√	Yes
27	H4N6F1	4	6	1	2053.7458	2053.533	×	No
28	H5N4F3	5	4	3	2101.7556	2101.571	√	No
29	H6N4F2	6	4	2	2117.7505	2117.567	√	No
30	H5N5F2	5	5	2	2158.7771	2158.541	√	No
31	H6N5F1	6	5	1	2174.7720	2174.591	+	No
32	H4N6F2	4	6	2	2199.8037	2199.585	√	No

33	H5N6F1	5	6	1	2215.7986	2215.570	×	No
34	H6N6F0	6	6	0	2231.7935	2231.558	+	No
35	H5N5F3	5	5	3	2304.8350	2304.575	√	No
36	H6N5F2	6	5	2	2320.8299	2320.604	×	No
37	H5N6F2	5	6	2	2361.8565	2361.589	√	No
38	H6N5F3	6	5	3	2466.8878	2466.621	√	No
39	H5N6F3	5	6	3	2507.9144	2507.661	√	Yes
40	H6N6F2	6	6	2	2523.9093	2523.572	×	No
41	H6N5F4	6	5	4	2612.9457	2613.742	√	No
42	H6N6F3	6	6	3	2669.9672	2670.694	×	No

4.4.3 Imaging of N-glycans in different regions of mouse brain sections

The ion images corresponding to 5 representative fucosylated N-glycans, H4N4F2, H5N4F3, H5N5F2, H4N6F2 and H6N5F4 are presented in Figure 4-4B-F. The signal intensity for each ion images is obtained by dividing the peak area of each glycan to the normalized peak area of the internal glycan standard (DP7) that had been spiked in PNGase F digestion solution during printing. In this study we have divided the brain into three major regions of brainstem (BS), isocortex (IsoCTX) and hippocampal formation (HPF). The AAL lectin histostaining of an adjacent mouse brain section is shown in Figure 4-4A. AAL preferentially binds to fucose (α-1,6) or (α-1,3) linked to N-acetylhexosamine. The AAL staining as well as the ion images indicate that fucosylation occurs in all regions of the brain, however, its prevalence seems to depend on the region. The AAL staining is strongest in the IsoCTX followed by HPF and generates the weakest signal in the BS (Figure 4-4A). Thirty fucosylated N-glycans are identified in this study, which comprises more than 70% of total number of glycans. This diversity in the number of the fucosylated N-glycans is also observed in their spatial distribution. While some of the fucosylated N-glycans such as H4N6F2 (Figure 4-4E) are more abundant in the BS, the majority of them show a stronger presence in the IsoCTX and HPF. In summary, the

fucosylation increases from the center of the brain towards the cortex. One crucial fact in comparing the lectin histostaining data with the MSI images is that the specificity of the lectins is far lower than the mass spectrometry. In fact, AAL staining depicts a superposition of all the (α-1,6) or (α-1,3) linked fucose-containing N- or O-linked glycans. Therefore, even though similar patterns between the two are expected, lectin histostaining results are not necessarily reflective of the distribution of single N-glycans over the tissue, which explains the differences observed between the lectin staining and MSI results for each individual glycan.

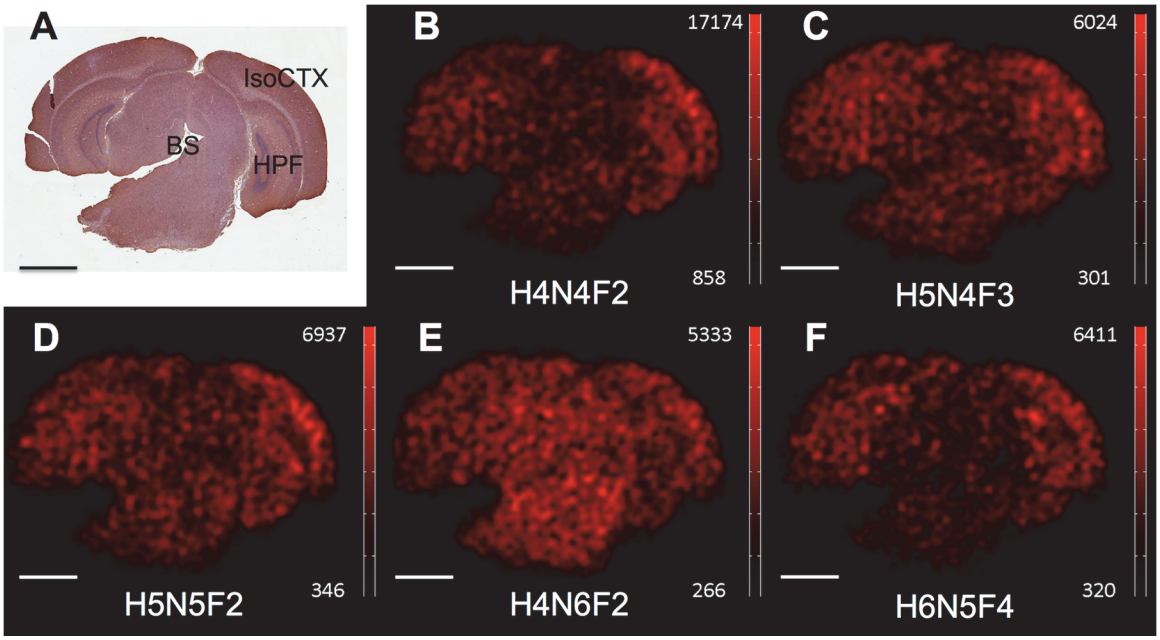


Figure 4-4. Ion images of representative fucosylated glycans along with AAL staining of an adjacent tissue section.
(A) Based on the AAL histostaining, fucosylation occurs in all regions of the brain, with a relatively higher abundance in the IsoCTX compared to the HPF, while BS has the lowest abundance of fucosylation. The ion images correspond to the peaks detected at (B) 1793.523 (H4N4F2), (C) 2101.571 (H5N4F3), (D) 2158.541 (H5N5F2), (E) 2199.585 (H4N6F2), and (F) 2613.742 (H6N5F4) Da. The signal intensity in each ion image is calculated by dividing the area of the corresponding peak by the normalized peak area of the internal glycan standard. Scale bar, 2 mm.

The ion images corresponding to 5 oligomannose structures of Man5 (H5N2F0), Man6 (H6N2F0), Man7 (H7N2F0), Man8 (H8N2F0) and Man9 (H9N2F0) are depicted in Figure 4-5B-F. The ConA lectin histostaining of an adjacent section is depicted in Figure 4-5A. ConA binds the α-mannose residues attached to the glycans. Therefore, both the oligomannose and hybrid N-glycans are potential targets for ConA. The ConA staining shows that terminal α-mannose residues are present in all of the aforementioned regions, however the signal is stronger in the IsoCTX and BS compared to the HPF (Figure 4-5A). The ion images also confirm that oligomannose N-glycans are more abundant in the BS. The ConA staining shows a slight asymmetry between the left and right half of the brain, particularly in the IsoCTX area. This asymmetry, which could be due to the tissue sectioning, was also observed on the ion images of oligomannose N-glycans. Two additional adjacent tissue sections were similarly analyzed to ensure the reproducibility of the results.

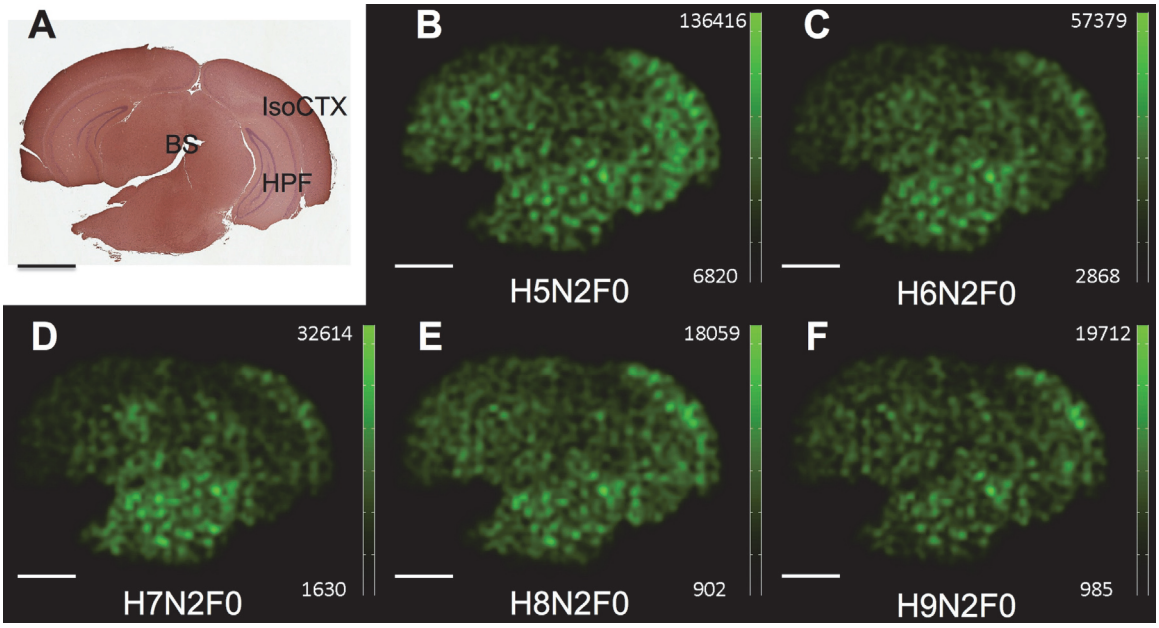


Figure 4-5. Ion images of representative oligomannose glycans along with ConA staining of an adjacent tissue section.

(A) Based on the ConA histostaining, oligomannose structures can be found in all regions of the brain. However, they seem to be more abundant in the IsoCTX and BS compared to the HPF. The ion images correspond to the peaks detected at (B) 1257.433 (Man5), (C) 1419.486 (Man6), (D) 1581.539 (Man7), (E) 1743.592 (Man8) and (F) 1905.645 (Man9). The signal intensity in each ion image is calculated by dividing the area of the corresponding peak by the normalized peak area of the internal glycan standard. Scale bar, 2 mm.

4.4.4 Imaging of N-glycans in glioblastoma tumor implanted in NOD/SCID mouse brain

Analysis of the tumor implanted mouse brain sections confirmed that glycosylation is altered during tumorigenesis. Mass spectrometry imaging of these sections revealed 13 N-glycans with different expressions levels in the tumor compared to the normal brain tissue. These N-glycans are listed in Table 4-2. To identify these glycans, two regions of interest (ROI) are defined in the brain, such that one ROI surrounds the tumor site and the other is mirrored in the normal half of the mouse brain. Two-sample t-test is applied to the signal intensities in the ROIs and a p-value threshold of 0.001 is used to determine which glycans differ in the tumor. Of the 13 altered N-glycans, 10 are increased in the tumor. While there is no obvious glycan structural pattern similarity between the glycans with altered abundances, the less abundant glycans are more fucosylated. For instance, the majority of the non-fucosylated N-glycans such as oligomannose and non-fucosylated complex structures are more abundant in the tumor. Although few fucosylated N-glycans are altered in the tumor cells to draw a general conclusion, this observation suggests that highly fucosylated glycans carrying antennary fucose residues might be down-regulated in tumor cells. Ion images corresponding to some of these glycans are depicted in Figure 4-6 as examples. The reproducibility of the results was confirmed by imaging an adjacent

tissue section using MSI. In this study, we have reported an MSI based technique for imaging of N-linked glycans released from immobilized glycoproteins on FFPE sections and demonstrated the application of this technique with two examples. We studied the spatial distribution of 42 N-glycans on mouse brain coronal sections and also imaged distinct N-glycans in patient-derived glioblastoma tumor cells implanted in a mouse brain. Similar techniques have been developed for direct profiling of tissue glycans such as on-surface enzymatic digestion of N-glycans followed by liquid chromatography-mass spectrometry [116]. However, this method does not preserve the spatial information concerning the distribution of different N-glycans. The developed MSI-based method provides a unique tool for high-throughput imaging of N-glycans from FFPE tissue sections, which distinguishes it from more conventional histostaining methods. It provides unique information regarding the spatial distribution of specific glycan structures over the tissue. This information, combined with histology, can provide potentially invaluable insight into the histopathology of many diseases. The acquired images from the C57BL/6 mouse brain sections suggested that the level of glycosylation and the type of N-glycans varies in different brain structures. Fucosylation was predominantly observed in the brain, such that more than 70% of all the identified glycans appeared to be fucosylated. The most prominent difference in brain N-glycan structures was observed between cerebral cortex and brainstem. While oligomannose and non-fucosylated complex structures were more abundant in the brainstem, fucosylated N-glycans showed overall higher signal in the cortex. Our results in the mice brain tumor model showed considerable differences between the N-glycosylation in tumor versus adjacent normal tissues. Low-abundance N-glycans in the tumor cells had higher levels

of fucosylation. On the other hand, high-abundance N-glycans in the tumor cells mostly consisted of oligomannose and non-fucosylated complex glycans. Knowing the spatial distribution of N-glycans in different brain structures or pathologies can shed light on the roles that glycosylation plays in mediating the brain functions.

Table 4-2. N-glycans and their relative abundance in tumor versus the surrounding normal brain tissue.

Thirteen N-glycans with differential expression in the tumor were identified, most of which were more abundant in the tumor cells. N-glycans that are less abundant in tumor have higher levels of fucosylation.

#	Symbol	No. of Fuc residues	Theoretical mass $[M+Na]^+$	Abundance in tumor
1	H4N3F0	0	1298.4497	Higher
2	H6N2F0	0	1419.4759	Higher
3	H4N3F1	1	1444.5076	Higher
4	H5N3F0	0	1460.5025	Higher
5	H7N2F0	0	1581.5287	Higher
6	H5N4F0	0	1663.5819	Higher
7	H3N5F1	1	1688.6136	Lower
8	H8N2F0	0	1743.5815	Higher
9	H4N4F2	2	1793.6449	Lower
10	H5N4F1	1	1809.6398	Higher
11	H9N2F0	0	1905.6343	Higher
12	H6N4F1	1	1971.6926	Higher
13	H4N5F2	2	1996.7243	Lower

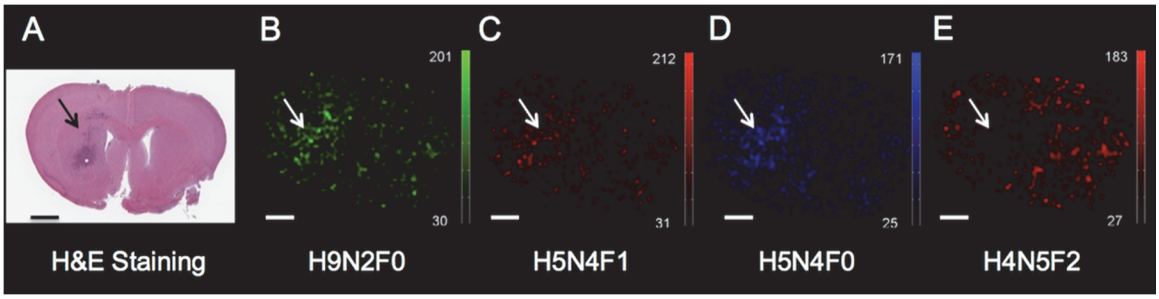


Figure 4-6. Ion images of tumor N-glycans along with H&E staining of an adjacent tissue section.
(A) Tumor cells are concentrated around the injection site. The ion images correspond to the peaks detected at (B) 1905.634 (H9N2F0), (C) 1809.640 (H5N4F1), (D) 1663.582 (H5N4F0), and (E) 1996.724 (H4N5F2) Da. The majority of the glycans are more abundant in the tumor; however, a few highly fucosylated glycan peaks are more abundant in the surrounding normal tissues. Scale bar, 1 mm.

In addition to FFPE sections, frozen tissue sections can also be analyzed by MALDI imaging [118]. For more elaborate results, the current imaging method can be combined with quantitative techniques using isotopic labeled standards for high accuracy quantitation. For example, by spiking in standard glycans labeled with stable isotope tags, one can relatively quantify the images acquired from the tissue sections for targeted glycans [65], [119]. In this study, we used the robotic application of the enzyme using the microarray printer with 100 µm spot-to-spot spatial resolution and the mass spectral images were acquired with spatial resolution of 100 µm. High-density deposition of enzyme, high resolution MS acquisition using a faster instrument with higher laser repetition rate [120] and imaging in the microscope mode using a triple focus time-of-flight mass spectrometer [121] could be used to increase the imaging spatial resolution. One of the other important limitations of MALDI-MS imaging is the complexity of the sample preparation. The dependency of the final results on the changes in the sample preparation makes the reproducibility challenging. Incorporation of automatic and semi-automatic sample handling can improve the reproducibility [122]–[125].

Current tissue analysis is mainly based on histological staining of nuclear and cytoplasm using H&E [126], [127], however molecular bases of the structural differences between cells can be further differentiated by antibodies to specific proteins using immunohistochemistry or nucleic acid probes to DNA or RNA. Immunohistochemistry methods are inherently low throughput and subsequently, tissue analysis using these methods operate based on single or few molecular probes. On the other hand, mass spectrometry provides simultaneous detection and identification of numerous targets, which sets it apart as a high-throughput technology. Lectin histostaining has shown that various cell types express distinct glycan patterns at different development stages. By revealing the glycosylation patterns at single cell level, mass spectrometry imaging of glycans can potentially be used to distinguish the cell type and development status based on its distinct glycan profile.

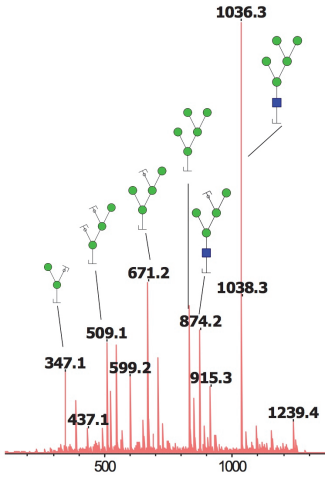
Glycan imaging offers some advantages over protein imaging for detection of disease biomarkers. Genetic mutations could result in synthesis of disease-specific aberrant glycan structures that are also cell specific. This combination provides a very specific target for identification of the underlying pathology, which can be detected using mass spectrometry imaging. These glycans are likely to modify different protein sequences, many of which contain multiple glycosylation sites. The presence and consequently detection of aberrant glycans on multiple glycosites and on various protein sequences using mass spectrometry amplifies the effect of the genetic mutation and therefore, compared to protein based assays, could potentially improve the detection sensitivity of the disease.

4.5 Appendix: Tandem mass spectrometry analysis of PNGase-F released glycan peaks

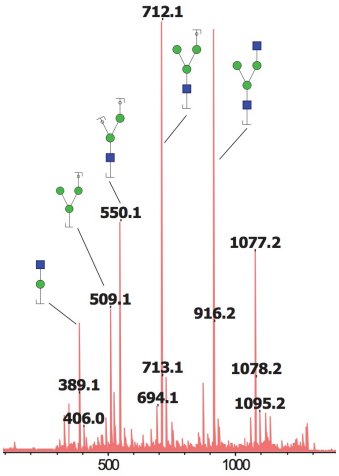
To confirm that the detected mass spectral peaks correspond to glycans, collision-induced dissociation tandem mass spectrometry is performed on PNGase-F released peaks on the mouse brain tissue sections. For the few glycans, where the precursor signal of the mass spectrometry imaging spectrum was too low for tandem mass spectrometry, brain tissue extracts are used to extract the brain glycans using the glycoprotein immobilization for glycan extraction (GIG) as described previously [26] and the N-glycans are analyzed by MALDI-MS for acquiring the MS/MS signal. Briefly, the proteins were extracted by sonicating the tissue in RIPA buffer, and were conjugated to AminoLink beads (Thermo Fisher Scientific Inc., Rockford, IL) by reductive amination. The glycans were released from the immobilized proteins by PNGase F, then purified in carbograph columns and finally analyzed by MALDI-MS. The tandem MS spectra that were acquired from the mouse brain glycan extracts are marked by an asterisk in Table 4-3. The fragment ions are assigned with the help of GlycoWorkbench software.

Table 4-3. Tandem spectra of PNGase F-released N-glycans.

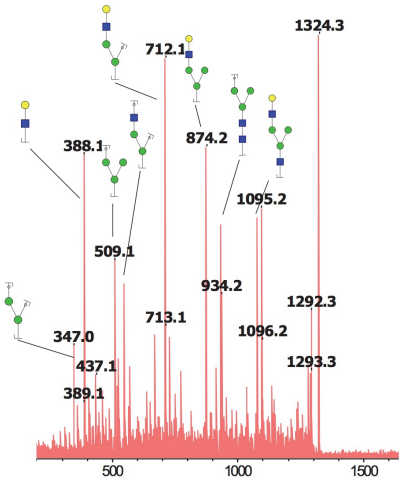
Tandem MS spectra are generated for all the detected glycan peaks to confirm their identity. The spectra were directly acquired from the tissue section for most of the peaks. In the cases, where the precursor peak was too low to generate decent MS/MS signal from the tissue section, mouse brain glycan extracts were used instead to acquire tandem MS spectra. These spectra are marked by an asterisk.



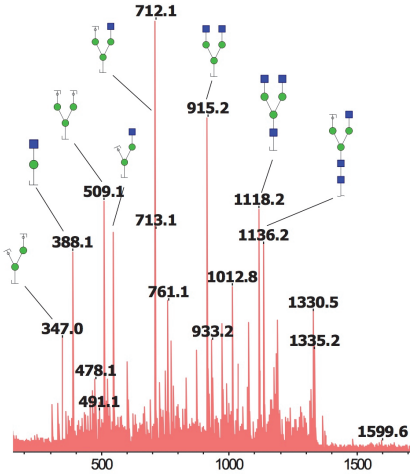
Precursor m/z 1257.4, glycan H5N2F0



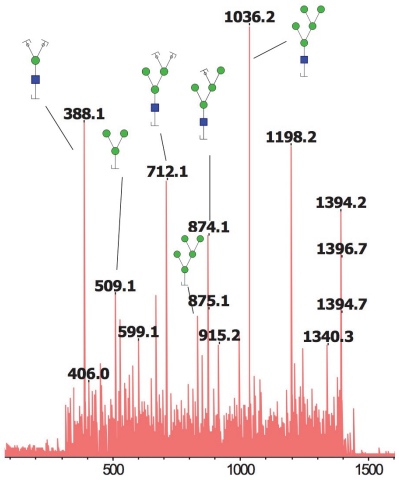
Precursor m/z 1282.5, glycan H3N3F1



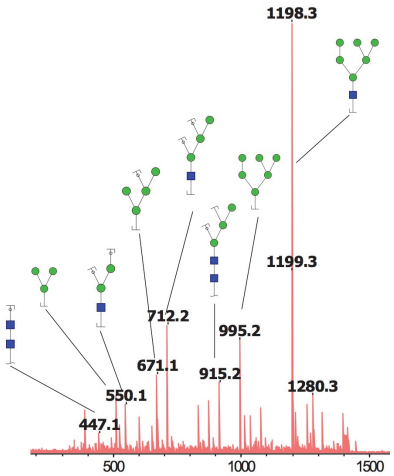
Precursor m/z 1298.4, glycan H4N3F0



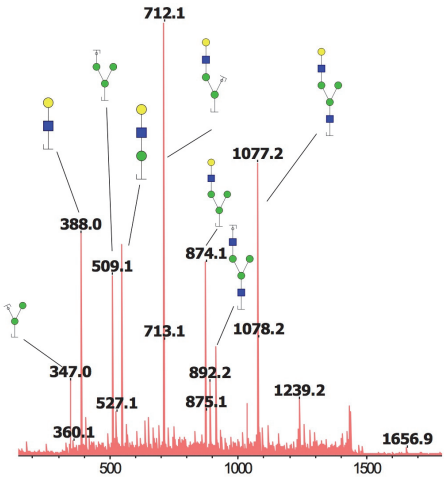
Precursor m/z 1339.5, glycan H3N4F0



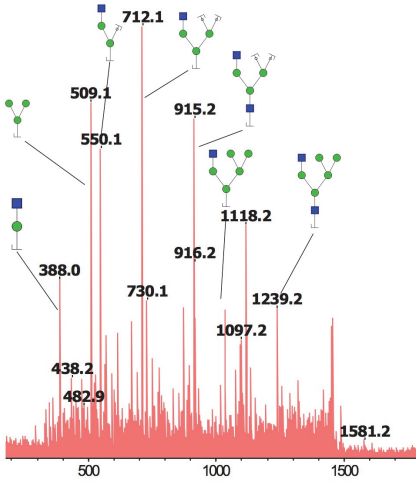
Precursor m/z 1403.5, glycan H5N2F1



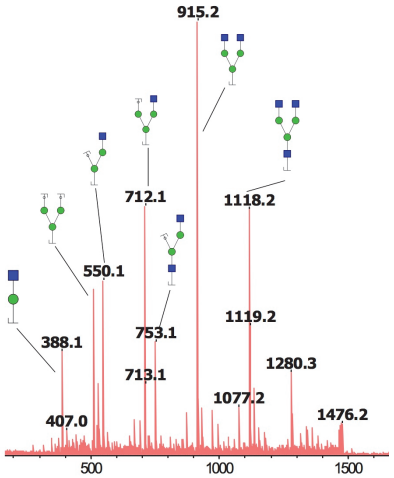
Precursor m/z 1419.5, glycan H6N2F0



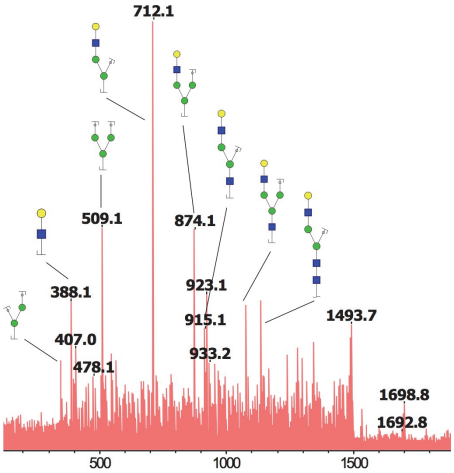
Precursor m/z 1444.5, glycan H4N3F1



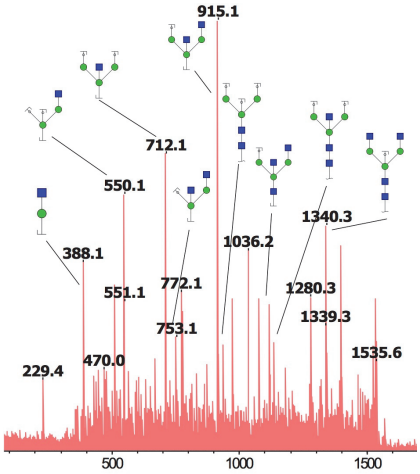
Precursor m/z 1460.5, glycan H5N3F0



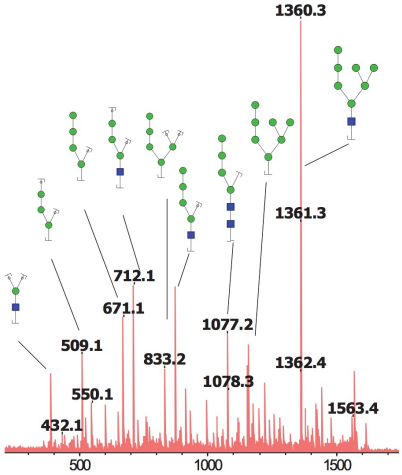
Precursor m/z 1485.5, glycan H3N4F1



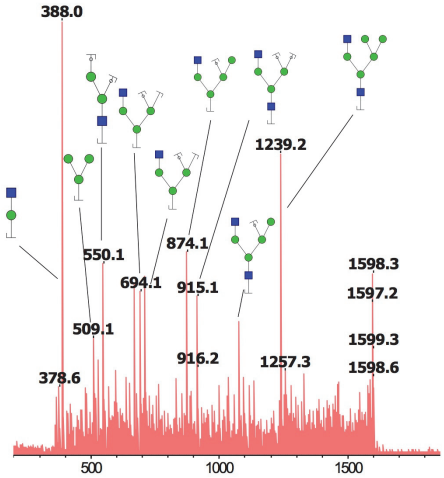
Precursor m/z 1501.5, glycan H4N4F0



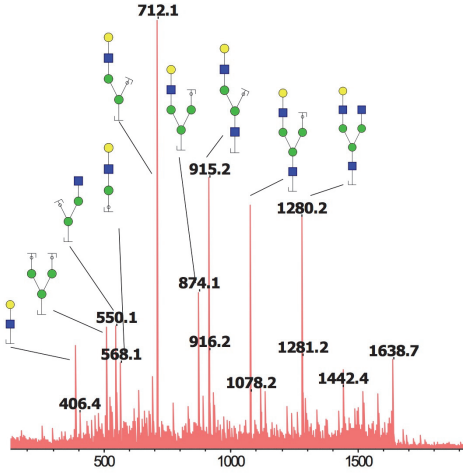
Precursor m/z 1542.6, glycan H3N5F0



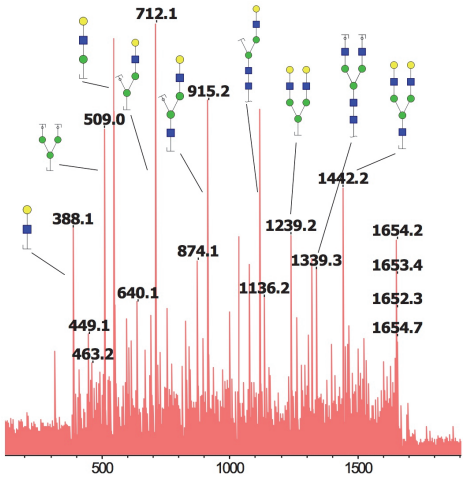
Precursor m/z 1581.5, glycan H7N2F0



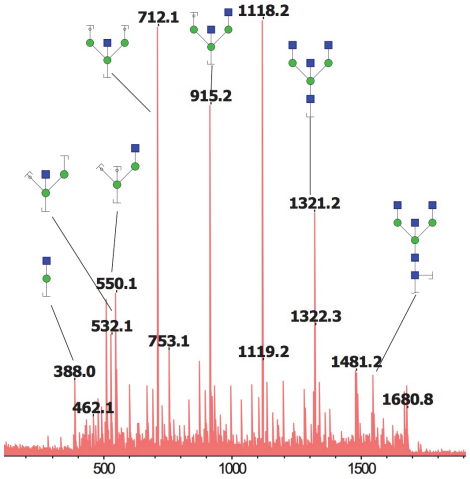
Precursor m/z 1606.6, glycan H5N3F1



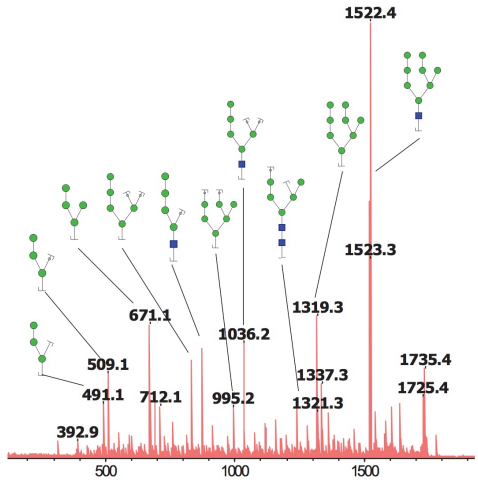
Precursor m/z 1647.6, glycan H4N4F1



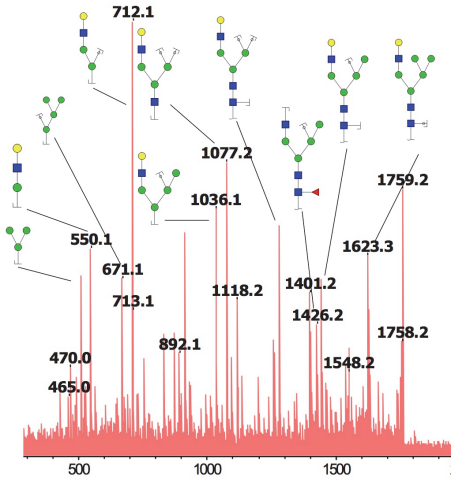
Precursor m/z 1663.6, glycan H5N4F0



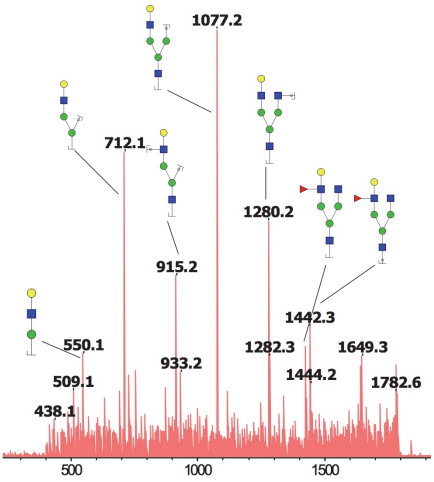
Precursor m/z 1688.6, glycan H3N5F1



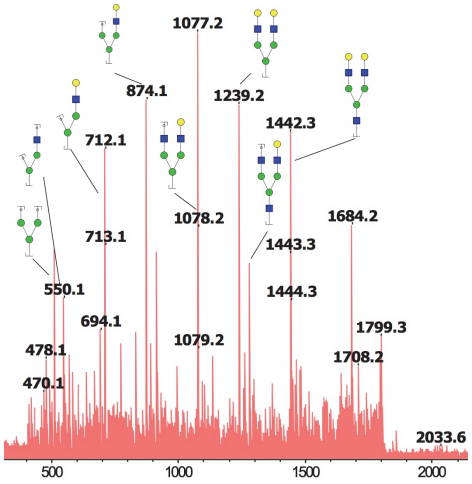
Precursor m/z 1743.6, glycan H8N2F0



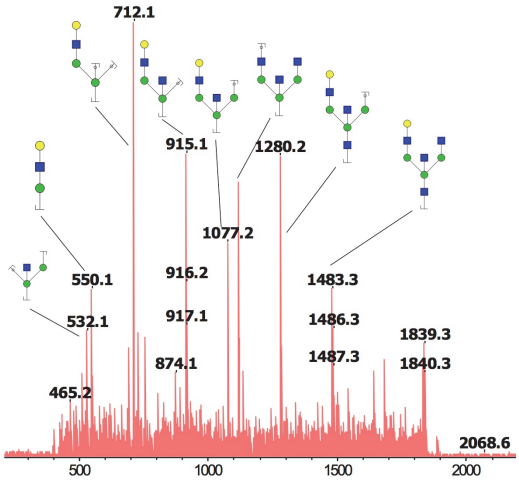
*Precursor m/z 1768.6, glycan H6N3F1

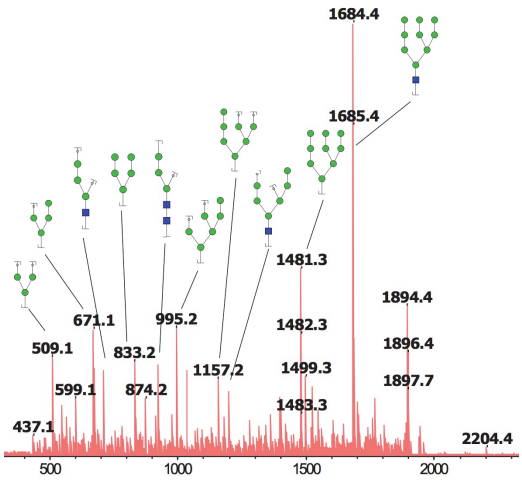


Precursor m/z 1793.6, glycan H4N4F2

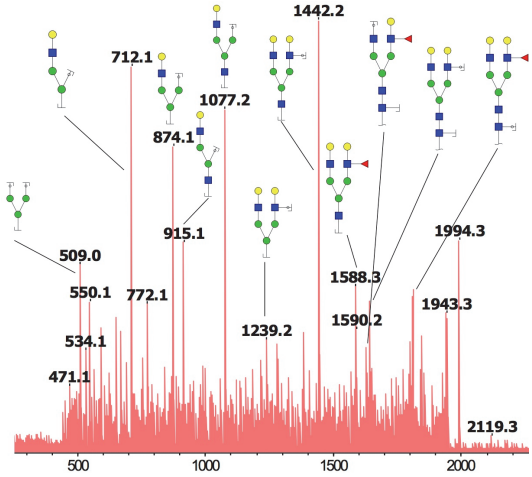


Precursor m/z 1809.6, glycan H5N4F1

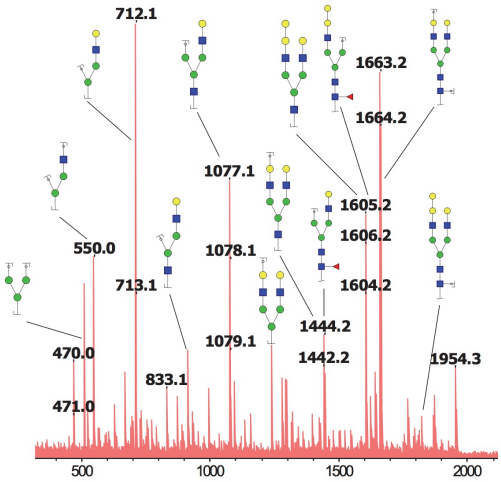




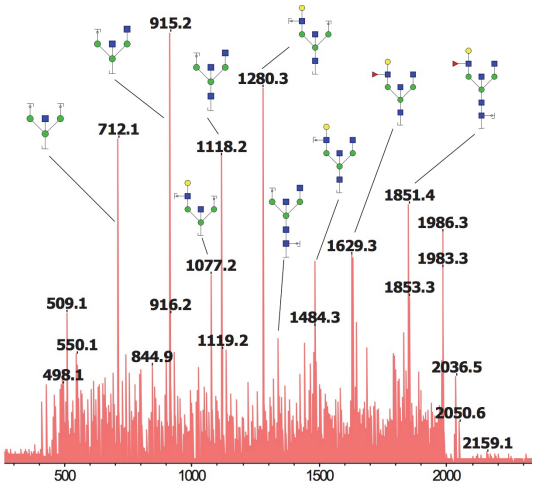
Precursor m/z 1850.7, glycan H4N5F1



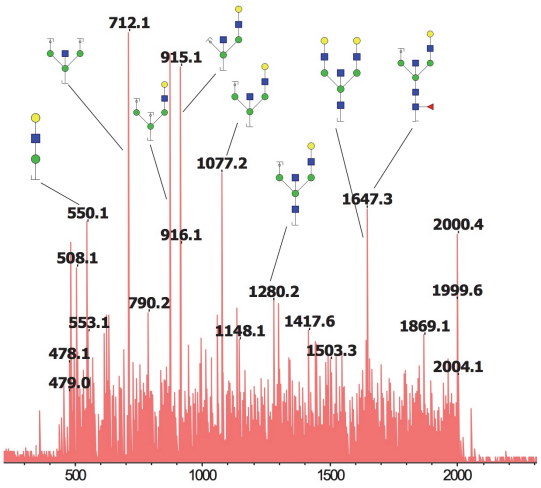
Precursor m/z 1905.6, glycan H9N2F0



Precursor m/z 1955.7, glycan H5N4F2

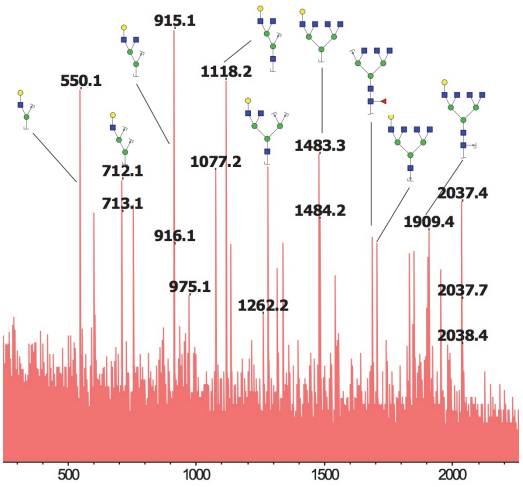


*Precursor m/z 1971.7, glycan H6N4F1

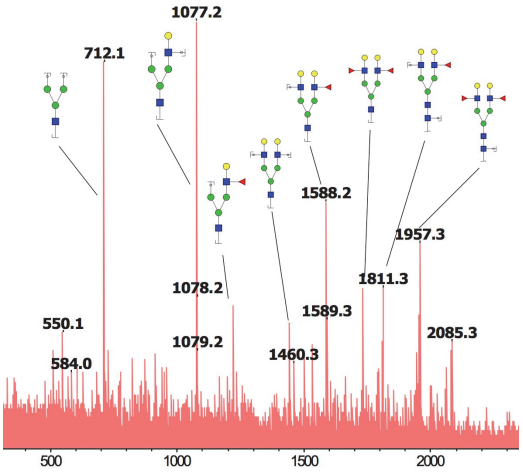


Precursor m/z 1996.7, glycan H4N5F2

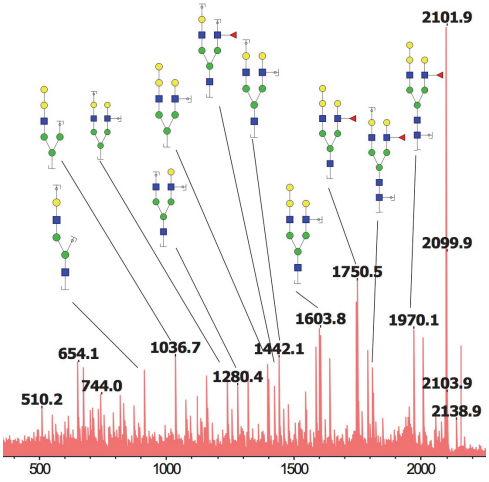
Precursor m/z 2012.7, glycan H5N5F1



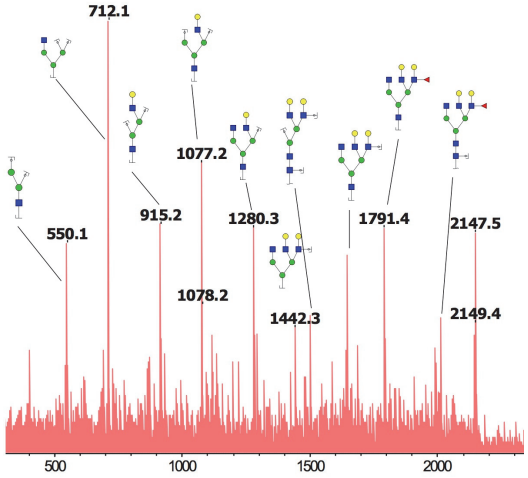
*Precursor m/z 2053.7, glycan H4N6F1



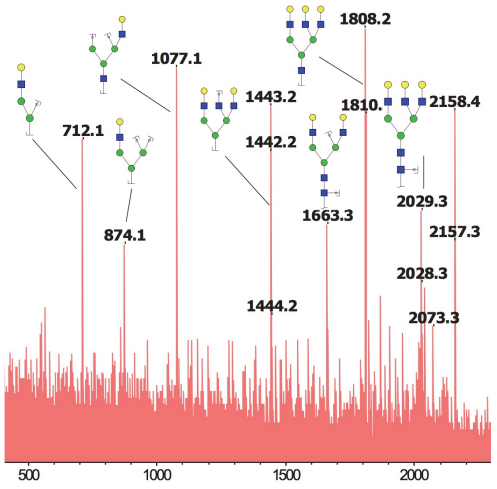
*Precursor m/z 2101.8, glycan H5N4F3



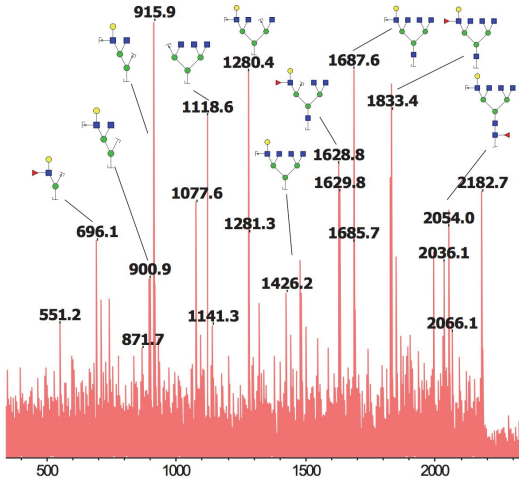
*Precursor m/z 2117.8, glycan H6N4F2



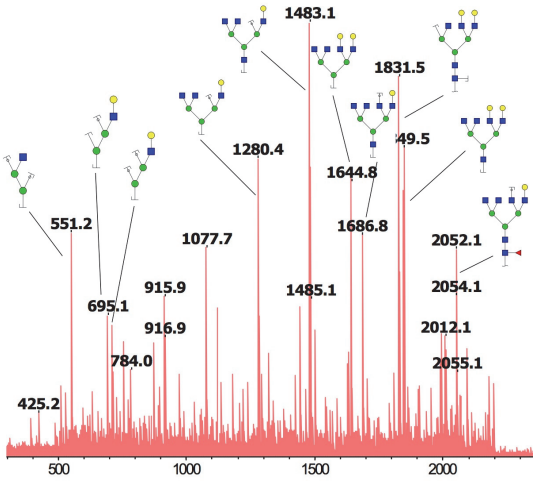
*Precursor m/z 2158.8, glycan H5N5F2



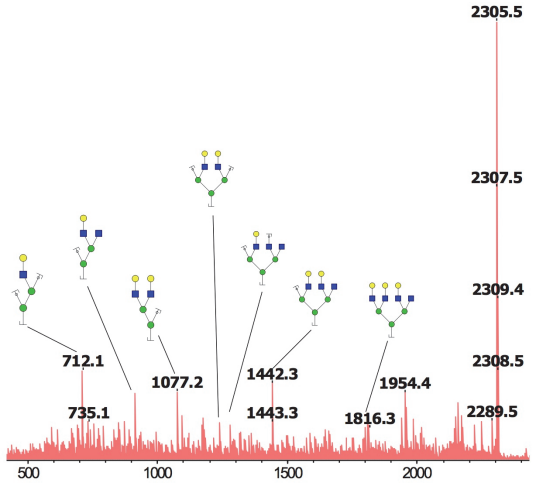
*Precursor m/z 2174.8, glycan H6N5F1



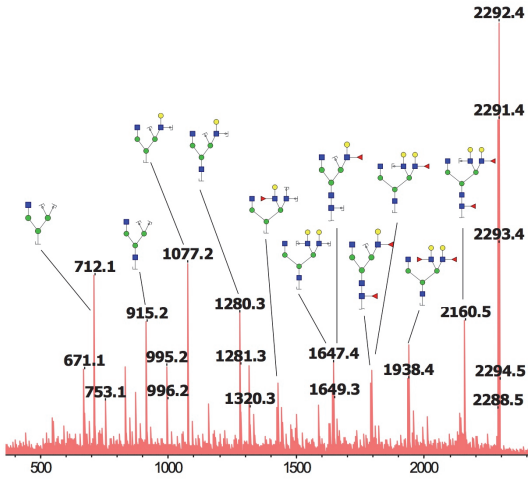
*Precursor m/z 2199.8, glycan H4N6F2



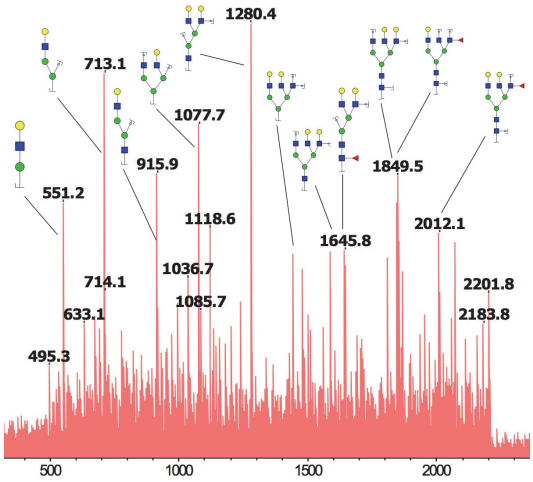
*Precursor m/z 2215.8, glycan H5N6F1



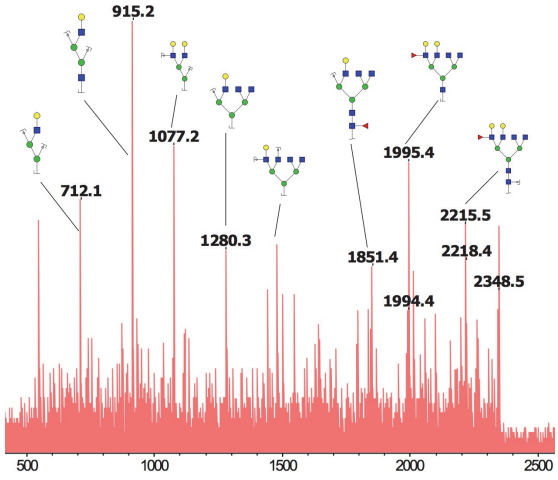
*Precursor m/z 2231.8, glycan H6N6F0



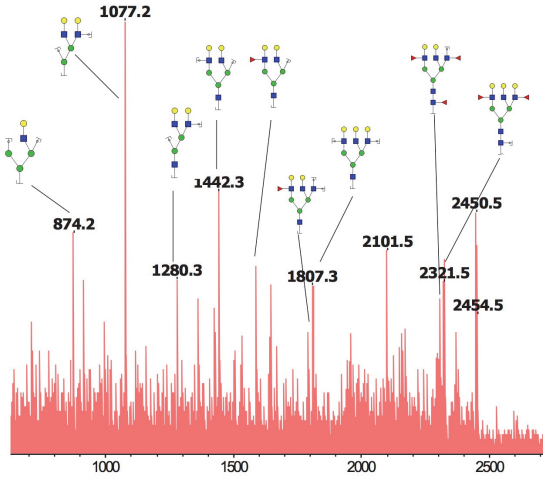
*Precursor m/z 2304.8, glycan H5N5F3



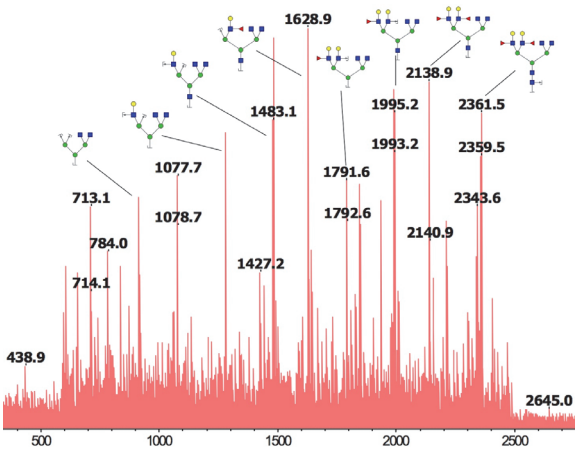
*Precursor m/z 2320.8, glycan H6N5F2



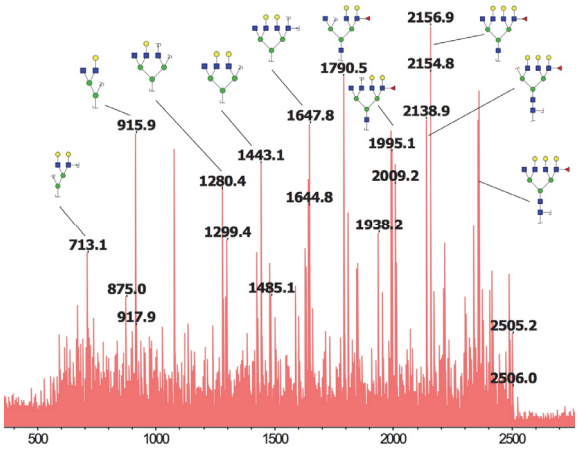
*Precursor m/z 2361.9, glycan H5N6F2



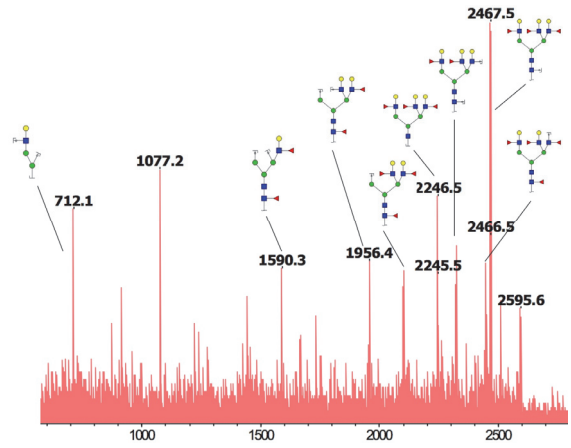
*Precursor m/z 2466.9, glycan H6N5F3



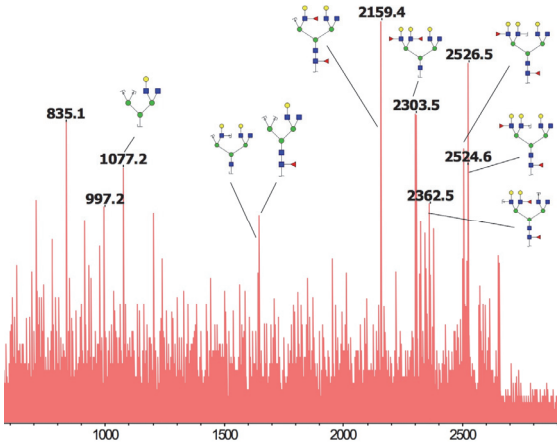
Precursor m/z 2507.9, glycan H5N6F3



*Precursor m/z 2523.9, glycan H6N6F2



*Precursor m/z 2612.9, glycan H6N5F4



*Precursor m/z 2670.0, glycan H6N6F3

Chapter 5. Spectral Library Matching for Site-Specific Assignment of Tandem Mass Spectra to Intact N-glycopeptides

5.1 Summary

Glycoprotein changes occur in not only protein abundance, but also the occupancy of each glycosylation site by different glycoforms during biological or pathological processes. Recent advances in mass spectrometry instrumentation and techniques have facilitated analysis of intact glycopeptides in complex biological samples by allowing the users to generate spectra of intact glycopeptides with glycans attached to each specific glycosylation site. However, assigning these spectra, leading to identification of the glycopeptides, is challenging. Here, we report a spectral library matching algorithm for site-specific identification of intact glycopeptides using higher-energy collisional dissociation (HCD) fragmentation of complex samples. In this algorithm, a spectral library of glycosite-containing peptides in the sample was built by analyzing the isolated glycosite-containing peptides using HCD LC-MS/MS. Spectra of intact glycopeptides were selected by using glycan oxonium ions as signature ions for glycopeptide spectra. These oxonium-ion containing spectra were then compared with the spectral library generated from glycosite-containing peptides, resulting in assignment of each intact glycopeptide tandem MS spectrum to a specific glycosite-containing peptide. The glycan occupying each glycosite was determined by matching the mass difference between the precursor ion of intact glycopeptide and the glycosite-containing peptide to a glycan database. We applied spectral library matching to analyze LC-MS/MS spectra of protein extracts from prostate tumor LNCaP cells. Without enrichment of glycopeptides from global tryptic peptides and at a false discovery rate of 1%, 1008 glycan-containing tandem MS spectra were assigned to 769 unique intact N-linked glycopeptides, representing 344 N-linked glycosites with 57 different N-glycans. The implemented

algorithm assigns the HCD LC-MS/MS generated spectra of intact glycopeptides in an automated and high throughput manner. Additionally, spectral library matching gives the user the possibility of identifying novel or modified glycans on specific glycosites that might be missing from the pre-determined glycan databases.

5.2 Introduction

Mass spectrometry analysis is routinely used for characterization of glycans and peptides in recombinant protein and complex biological samples [63], [107], [128]. Various fragmentation techniques have been investigated for analysis of glycopeptides. These techniques, mainly operating on the basis of vibrational or electronic excitation energies, yield unique fragmentation patterns [85]. For example, collision-induced dissociation (CID) results in fragmentation of the attached glycan leaving the peptide backbone intact, while electron transfer dissociation (ETD) breaks the peptide backbone leaving the intact glycans attached to the amino acids, thus revealing the glycosylation site [74], [88], [89]. Therefore, multiple tandem approaches are usually required to fully characterize the structure of intact glycopeptides [89]. Higher-energy collisional dissociation (HCD) is another fragmentation method that results in fractionation of both glycans and peptides of glycopeptides [90]. Eliminating the need for multiple tandem fragmentation approaches, HCD fragmentation provides extensive information regarding the peptide backbone of each intact glycopeptide, including b and y ions of peptide as well as the mass of glycan attached to the glycopeptide [84], [85]. However, assignments of tandem mass spectra to intact glycopeptides, including characterization of specific glycosites and their occupying glycans, is a challenging but critical step to determine the microheterogeneity of the glycosylation at each glycosite [80], [84]–[86], [91], [99], [129]–[131].

In this study, we present a novel algorithm for automatic identification of intact glycopeptides from HCD spectra of complex biological samples based on spectral library matching. This algorithm uses the spectral library built from proteomics analysis of glycosite-containing peptides from the complex biological sample using HCD. It takes advantage of the fact that tandem mass spectra of intact glycopeptides contain oxonium ions and the fragment ions generated through HCD fragmentation of intact glycopeptides resemble the pattern of same ions in the tandem MS spectra of their deglycosylated counterparts. Therefore, the tandem MS spectra of HCD-fragmented intact glycopeptides are selected using oxonium ions and matched against the generated spectral library to identify the glycosite-containing peptides and the glycosites. After identification of the peptide portion of the intact glycopeptide, the exact precursor mass is used to identify the glycan portion by matching with a glycan database. In this study, we analyzed protein lysates from LNCaP cells and identified 769 unique intact N-glycopeptides. Besides the glycopeptide-matching tandem MS spectra, we identified other tandem MS spectra that matched to glycosites in the spectral library, while their corresponding glycan masses were not included in the glycan database. Further investigation of the unknown glycans, led to assignment of glycans that were modified by an additional moiety with a mass of 17.018 Da. This modification was particularly shown for high-mannose structures. Identifying the peptide portion of the intact glycopeptide first, spectral library matching allows for identification of possibly novel glycans as well as modified glycans attached to the peptides, which could be of great value particularly in recognizing the pathologically induced changes of glycosylation.

5.3 Methods

5.3.1 Sample preparation

5.3.1.1 Materials and reagents

LNCaP cell line was obtained from ATCC. Hydrazide beads were purchased from Bio-Rad laboratories (Hercules, CA). Peptide-N-Glycosidase F (PNGase F) was from New England Biolabs (Ipswich, MA). Sequencing-grade trypsin was purchased from Promega (Madison, WI). C18 desalting cartridges were purchased from Waters (Milford, MA). All other reagents were purchased from Sigma Aldrich (St. Louis, MO) unless otherwise specified.

5.3.1.2 LNCaP sample preparation and mass spectrometry analysis

LNCaP cells were lysed and the extracted proteins were alkylated by treating with iodoacetamide. The sample was digested using trypsin at 37°C overnight. Tryptic peptides were labeled with iTRAQ reagents according to the manufacturer instructions, then desalted and purified using C18 columns. Ninety percent of the sample was enriched for glycosite-containing peptides using the solid phase extraction of glycosite-containing peptides (SPEG) method. Briefly, the samples were treated with a 10 mM sodium periodate solution and conjugated to hydrazide beads at room temperature in the dark on a shaker. Non-specific binding was removed by washing the hydrazide beads. The glycosite-containing peptides were detached from the immobilized N-glycans with PNGase F treatment and collected for mass spectral analysis. The remaining 10% of the tryptic peptide mixture was dried in a speedVac, resuspended in 0.4% acetic acid and fractionated for mass spectral analysis. The global tryptic peptides were fractionated using basic reverse phase liquid chromatography (bRPLC). The collected 96 fractions were combined into 24 fractions. The glycosite-containing peptides that were isolated

from the cells using the SPEG technique were directly analysed by LC-MS/MS without fractionation. A 1 μg of aliquot of each sample was separated through a C18 column on a Dionex Ultimate 3000 RSLC nano system (Thermo Scientific) and analysed on a Q Exactive mass spectrometer (Thermo Scientific). Data-dependent HCD fragmentation was performed on the 15 most abundant ions using an isolation window of 4 *m/z*. Using charge state screening, unassigned, singly, eight and more than eight protonated ions were rejected. In addition, an exclusion 25 second window was applied to avoid multiple selections of the same ions.

5.3.2 Data analysis

5.3.2.1 Building the experimental spectral library (ESL) for glycosite-containing peptides

The mass spectrometry results of the SPEG-enriched glycosite-containing peptides were analysed using SEQUEST in the Proteome Discoverer software with the following parameters: fixed Cys modification, dynamic PNGase-F facilitated conversion of Asn to Asp, and dynamic oxidation of Met in addition to a maximum of two miscleavages. The top peptide match for each spectrum was selected. For each peptide, list of singly, doubly and triply charged b and y fragment ions were generated. The tandem MS spectra matched to each peptide were searched for the presence of these ions. The experimental spectral library, which contained the list of target glycosite-containing peptides and their present fragment ions, was built based on the results of this search. Any peptide with less than 4 observed fragment ions was removed from the target database and the experimental spectral library.

5.3.2.2 Preprocessing

The generated raw files containing the acquired mass spectra were converted to mzXML files using the msconvert utility in the Trans-Proteomic Pipeline software. The 'centroid all scans' option was selected. The mzXML file corresponding to each of the tryptic global peptide run was opened in MATLAB. The tandem MS spectra of the glycopeptides were distinguished from peptide tandem MS based on the presence of oxonium ions. These ions belong to glycan free monosaccharides or disaccharides that were fragmented during the tandem mass spectrometry analysis. In this step, the tandem MS spectra including at least two of the oxonium ions with the masses of 138 (internal fragment of HexNAc), 145 (Hex – H_2O), 163 (Hex), 168 (HexNAc - $2H_2O$), 186 (HexNAc - H_2O), 204 (HexNAc), 325 (Hex_2), 366 (HexHexNAc), 274 (Neu5Ac - H_2O), or 292 (Neu5Ac) were isolated as oxonium-ion containing spectra. For the spectra with more than 100 peaks, oxonium ions were searched in the top 10% of the mass spectral peaks within a 10 ppm window.

5.3.2.3 Matching the spectra of HCD-fragmented glycopeptides with the ESL

Each oxonium ion-containing tandem MS spectrum was compared with the compiled experimental spectral library (ESL). For each target peptide in the ESL, the percentage of the b and y ions that were observed in the tandem MS spectrum was calculated. In addition, a list of candidate intact peptide ions was generated for each ESL peptide and the tandem MS spectrum was searched for the presence of these ions. A total of 9 intact peptide ions were considered including singly, doubly and triply charged intact peptide and intact peptide + HexNAc and singly charged intact peptide + $HexNAc^{0,2}$ cross-ring cleavage ion, intact peptide + FucHexNAc and intact peptide + $HexNAc_2$. The peptide

matches were first filtered based on the number of their observed intact peptide ions. The number of required ions for each peptide depends on the length of the peptide and is shown in Table 5-1. The results were further refined by applying an FDR of 1%. A 50 ppm window was used for matching the b, y and intact peptide ions.

5.3.2.4 Assignment of glycans attached to glycosite-containing peptides at each glycosite

To identify the monoisotopic peak for each dissociated ion and correct the mass shift, the MS spectra were averaged over a window of 15 spectra centered at the precursor MS. The first peak in the averaged isotopic cluster corresponding to the precursor mass was picked as the monoisotopic peak. The glycan composition was then deduced by first calculating its mass from the monoisotopic mass and the peptide mass and then running an exact match search against the glycan database with a mass tolerance of 10 ppm.

5.4 Results

5.4.1 Building the spectral library for glycosite-containing peptides

The application of spectral library matching requires an experimental spectral library (ESL) of glycosite-containing peptides as a basis to identify the peptide portion of the intact glycopeptide in each tandem MS spectrum. The ESL can be generated from the glycosite-containing peptides isolated from the sample. The ESL contains the experimentally observed b and y ions for each of these peptides, which are the most dominant ions in the HCD spectra compared to other ion types. To generate the ESL, the glycosite-containing peptides were first isolated from the sample using the solid-phase extraction of N-linked glycosite-containing peptides (SPEG) technique [61], [63]. The

mass spectral results of the SPEG-enriched glycosite-containing peptides were searched using SEQUEST in the Proteome Discoverer (PD) software to identify 2,213 N-linked glycosite-containing peptides in the sample and the PD output was used to compile the list of peptides in the ESL (Supplementary Table 5-1). The list of b and y fragment ions, including doubly and triply charged ions, was generated for the ESL peptides. The spectral library was built by identifying all the experimentally observed b and y ions for each identified peptide in the sample.

5.4.2 Matching the spectra of HCD-fragmented glycopeptides with the ESL

Understanding the pattern of HCD fragmentations of glycopeptides is crucial in identifying optimal algorithms for matching of the spectra to glycopeptides. Based on observation of HCD fragmented glycopeptide tandem MS spectra, the most abundant ions are classified into four main groups of 1) oxonium ions, 2) peptide b and y ions, 3) intact peptide attached to partial glycan ions and 4) peptide b and y ions attached to partial glycan ions. In the majority of the tandem MS spectra corresponding to glycopeptides, a minimum of two oxonium ions are observed among the highest mass spectral peaks. We used this characteristic to select glycopeptide tandem MS spectra based on the presence of at least 2 signature oxonium ions in the highest 10% of the peaks. The second group of ions generated by HCD fragmentation of glycopeptides is the peptide b and y ions. These ions, which lie in the mass range between the first and the third group, are not as abundant as oxonium ions. However, they have a similar pattern to the fragmentation pattern of their deglycosylated glycosite-containing peptide counterparts. Figure 5-1A shows the tandem MS spectra of the deglycosylated ‘YHYN#GTFEDGK’ peptide, while the tandem MS spectrum corresponding to the

glycosylated ‘YHYN#GTFEDGK’ is depicted in Figure 5-1B. Comparing the two spectra showed that although the overall signal intensity of b and y ions was lower in the fragmented glycopeptide, the patterns of these fragment ions were similar between the spectra of fragmented intact glycopeptide and the deglycosylated glycosite-containing peptide that had been released from glycans using the SPEG technique. In the second step of the algorithm, we took advantage of this repeated pattern to identify the peptide portion of the intact glycopeptide for each tandem MS spectrum by evaluating its matching with the ESL. The overlap between each tandem MS spectrum of intact glycopeptide and each peptide entry in the ESL was estimated by calculating the percentage of the experimental b and y ions that the two share. The results were later refined by removing the matches whose overlap with the ESL did not reach a certain threshold. This threshold was determined by the desired false discovery rate (FDR). The third group of glycopeptide tandem MS ions of interest, i.e. the intact peptides with partial glycan structures were again among the highest peaks in the mass spectra. The ions of intact peptides and intact peptides attached to one or two HexNAc residues repeatedly were reported among the highest peaks [85], [90]. Therefore, the presence of peptide ions with or without partial glycan attachments was used as an additional criterion to identify the peptide portion of the glycopeptides. The forth group of ions i.e., peptide b and y ions attached to partial glycans can be used to further assign the remaining peaks to the glycopeptide ions. It should be noted that the presence of peptide b and y fragment ions as well as intact peptides with partial glycan ions depends on the length of the peptide, meaning that for longer peptides, b and y ions of peptide backbone are more dominantly observed, whereas for shorter peptides, intact peptide ions are more

prevalent. Therefore, a combination of all these ions is necessary to ensure the accuracy of the matching process. To account for the difference in peptide length, a peptide length-dependent threshold on the minimum number of present intact peptide ions was used to further refine the matches. The threshold was pre-specified for each peptide based on its length as shown in Table 5-1. At the end, by subtracting the mass of the identified peptide portion from the corrected precursor mass, we calculated the mass of the glycan portion of the glycopeptides. Figure 5-2 depicts the schematic workflow of the spectral library matching approach.

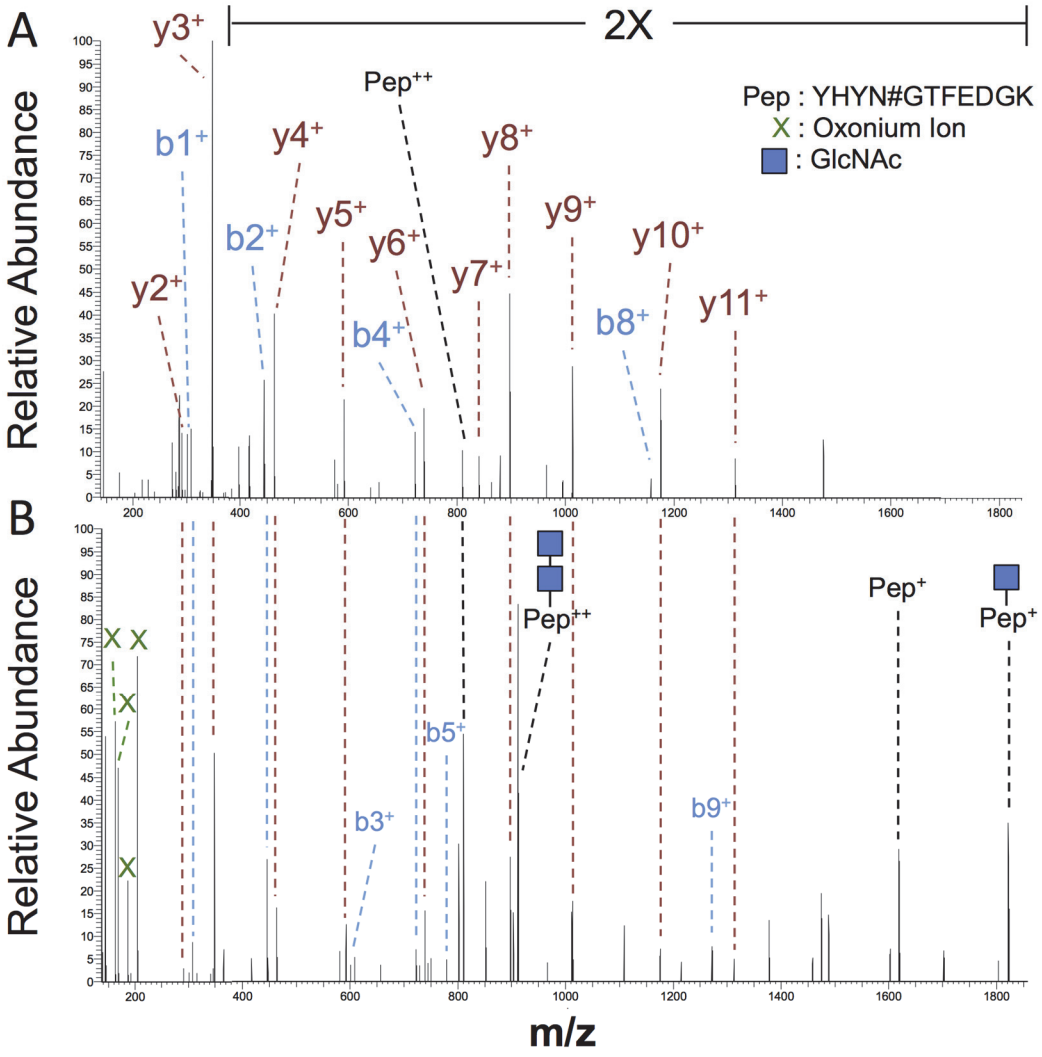


Figure 5-1. Comparison of the tandem MS spectra of HCD-fragmented glycosylated peptides with and without PNGase F treatment.

(A) MS/ MS spectrum of the deglycosylated glycosite-containing peptide "YHYN#GTFEDGK" is dominated by b and y fragment ions. (B) Tandem MS spectrum of a sample glycosylated peptide can be distinguished from the non-glycosylated peptides based on the presence of numerous oxonium ions marked by a cross. The intact peptide ions "YHYN#GTFEDGK" with partial glycan attachments are usually, and particularly for shorter peptides, the second most dominant set of ions in the tandem MS spectrum of a glycosylated peptide. The b and y fragment ions lie between the two aforementioned sets of ions and follow the pattern of the PNGase F-treated peptide.

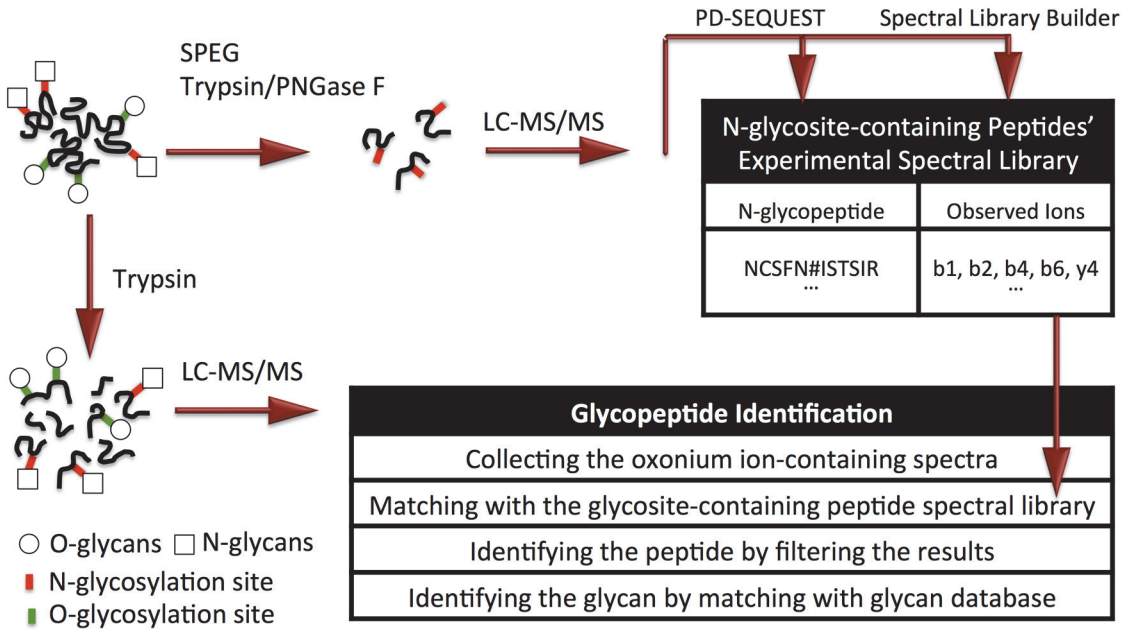


Figure 5-2. Schematic workflow of the spectral library matching approach.
The glycosite-containing peptides were isolated using the SPEG technique and analyzed with HCD LC-MS/MS analysis. The tandem MS spectra were assigned to glycosite-containing peptides by proteomics tools, such as PD-SEQUEST, and were used to build the corresponding sample-specific ESL by searching the spectra for b and y ions of each identified peptide. The ESL was then compared with the mass spectra of intact glycopeptides in the sample to identify the peptide portion of the intact glycopeptide corresponding to each oxonium-ion-containing tandem MS spectrum. The glycan portion was identified by matching its corresponding mass to the glycan database.

Table 5-1. The minimum number of required intact peptide ions and intact peptide ions with partial glycans for each glycosite-containing peptide based on its length.
The number of observed intact peptide ions decreases for longer peptides. Therefore, a length-dependent threshold on the number of required intact peptide ions is used to refine the match results.

Peptide Length	Minimum number of required intact peptide ions
5 < Peptide Length < 11	3
10 < Peptide Length < 16	2
15 < Peptide Length < 21	1
20 < Peptide Length	0

5.4.3 Estimation of the false discovery rate using decoy strategy

False discovery rate (FDR) is a crucial parameter in defining the specificity of the identification. The FDR can be calculated through either diversifying the spectra against the database of interest or diversifying the peptide database by adding decoy peptides to it. For decoy peptides, creating a reverse database is the most common method to generate a decoy database, because the reverse database resembles the target peptides in terms of number of peptides, peptide length and precursor molecular weight. However, the theoretical tandem MS spectra of the reverse database do not match the target database. Hence, matches to the reverse database resemble the random matches in the target database [132]. The algorithm, matching the ESL to the observed tandem spectra, not only used the b and y ions, but also the intact peptide ions with or without attached partial monosaccharides to narrow down the peptide-spectral matches (PSM). The theoretical tandem MS of the reverse database contains identical intact peptide ions to those of the target database, thus elevating the number of random and false matches. In this study, to generate the decoy database, we combined the amino acids from all SPEG-identified glycosite-containing peptides, shuffled them, and broke them into decoy peptide sequences with the same length as the target database. Figure 5-3 shows the mass distribution of the target database and an average of 10 randomly-generated decoy databases created by this strategy and depicts the similarity between these databases.

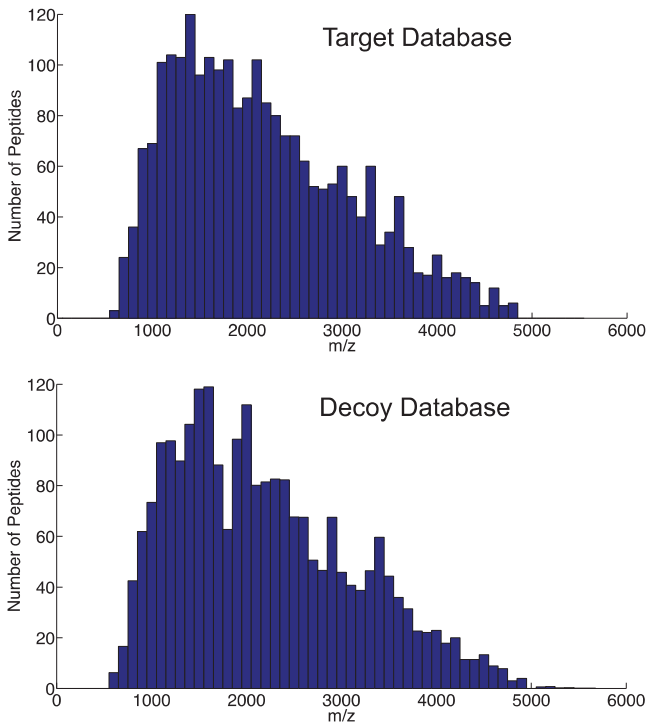


Figure 5-3. Comparison of the distribution of mass/charge (m/z) ratio between the target and decoy databases.
The decoy database was generated by shuffling the amino acids of the peptides in the target database and dividing them into peptides with the same lengths as the target database. The mass/charge ratio distribution of the decoy database resembles that of the target database, which is important for evaluating the FDR of the algorithm according to the number of false matches to the decoy database.

5.4.4 Assignment of glycans attached to glycosite-containing peptides at each glycosite

Correct assessment of the precursor mass is of great significance while assigning the glycan portion of the glycopeptide structure corresponding to each tandem MS spectrum. The abundance of glycopeptides is considerably lower than that of the peptides in a complex sample and the isotopic distribution of glycopeptides is different from that of peptides without glycans attached. This results in deviations in the isotopic pattern of glycopeptides such as obscuring the monoisotopic peak and subsequently increasing the possibility of inaccurate or wrong assignment of monoisotopic peaks of glycopeptides. Therefore, the precursor ion mass provided in the mzXML file might be as much as a few Daltons off. Using a shifted precursor mass will result in either match failure or matching to a wrong glycopeptide structure. For example, the mass difference between two fucose

residues and one N-acetylneuraminic acid (Neu5Ac) residue is equal to 1.02 Da. Therefore, an error in the detection of the right mass of monoisotopic peak in the glycopeptide ion cluster could result in assigning the wrong glycan structure to the tandem MS spectrum.

To determine the accurate mass of the monoisotopic peak of a glycopeptide, we calculated the average spectrum of consecutive MS spectra in the vicinity of the glycopeptide of interest over the elution time. The averaging improved the cluster isotopic pattern and the identification of the monoisotopic peak mass, which consequently improved the assignment of the glycopeptide spectrum. Figure 5-4A shows a glycopeptide isotopic cluster observed in a single MS1 spectrum and the precursor mass reported by the instrument software in the mzXML file. The algorithm identified the correct monoisotopic peak after averaging the spectra over a ~1-minute elution time window (Figure 5-4B).

After assignment of precursor mass and the peptide portion of the intact glycopeptide and glycosite, the exact mass of the glycan portion was determined by subtracting the peptide mass from the glycopeptide mass. To determine the glycan structure on each glycosite, the calculated glycan mass was compared with a glycan database and the glycan composition was determined at a mass tolerance of 10 ppm. The glycan database was composed by compiling several human serum or plasma N-glycan libraries [114], [115] with N-glycans identified from various human samples analyzed in our lab [26], [65]. All glycans with equal number of Hex, HexNAc, Fuc, and Neu5Ac monosaccharide residues in their compositions were grouped together and represented as a single entry, leading to 208 unique N-glycan compositions in the database (Supplementary Table 5-2).

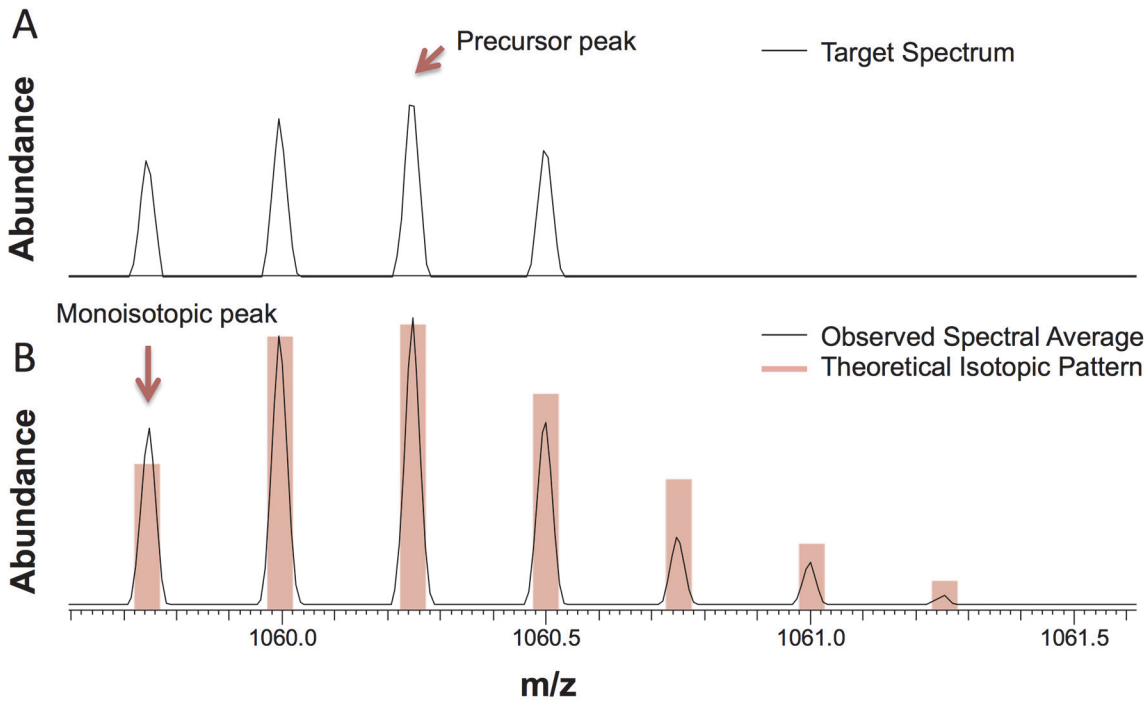


Figure 5-4. Detection of the glycopeptide monoisotopic peak.
(A) Isotopic pattern of a glycopeptide peak (TN#ITLVCKPGDLESAPVLR, Man9). The reported precursor mass is 2 Da off the monoisotopic peak. (B) Precursor mass correction by averaging a window of MS spectra over the glycopeptide elution time greatly improves the isotopic distribution of the cluster by comparison with the theoretical pattern, thus improving the detection of monoisotopic mass. The precursor peak and the monoisotopic peak are marked by a red arrow.

5.4.5 Glycoproteomics analysis of the LNCaP cells using spectral library matching

To identify glycopeptides from a complex sample, twenty-four fractions of LNCaP tryptic peptides and SPEG-enriched glycosite-containing peptides were analyzed, and glycopeptides were identified using the spectral library matching algorithm. The generated ESL for the LNCaP samples, which was built based on the mass spectrometric analysis of SPEG-enriched glycosite-containing peptides, contained 2,213 target peptides (Supplementary Table 5-1). Of the total number of 985,509 spectra in all 24 fractions, 7,243 contained at least 2 oxonium ions and were isolated as tandem spectra corresponding to glycopeptides.

With no filtering on the percentage of observed b and y ions, a total of 137,227 PSMs were attained, where each matched tandem MS scan was matched to an average of 23.9 PSMs. Refining the results based on a threshold on the minimum percentage of overlap between the ESL and the spectra to achieve a reasonable FDR, as expected, decreased the number of PSMs.

For estimation of the FDR, the decoy database was built as described and merged with the target database resulting in a total of 4,426 peptides in the peptide database. The FDR was calculated based on the percentage of PSMs matched to the decoy database. A curve was calculated by changing the threshold on the percentage of b and y ion overlap with ESL and calculating and plotting the FDR as a function of this threshold. The optimal threshold for refining the results was determined by the desired FDR on this curve. Figure 5-5 shows the FDR as a function of this threshold in red and black, where the red curve corresponds to spectral library matching and the black curve corresponds to similar analysis, with the only difference that the match refinement step based on the number of observed intact peptide ions was omitted. From this figure, we estimated that a threshold of 40% results in an FDR of approximately 1%. In addition, this figure demonstrated how the use of intact peptide ions for filtering the results improved the FDR. Using a 1% FDR cut down the number of PSMs to 4,213 and the average number of PSMs per matched tandem MS scan to 1.6.

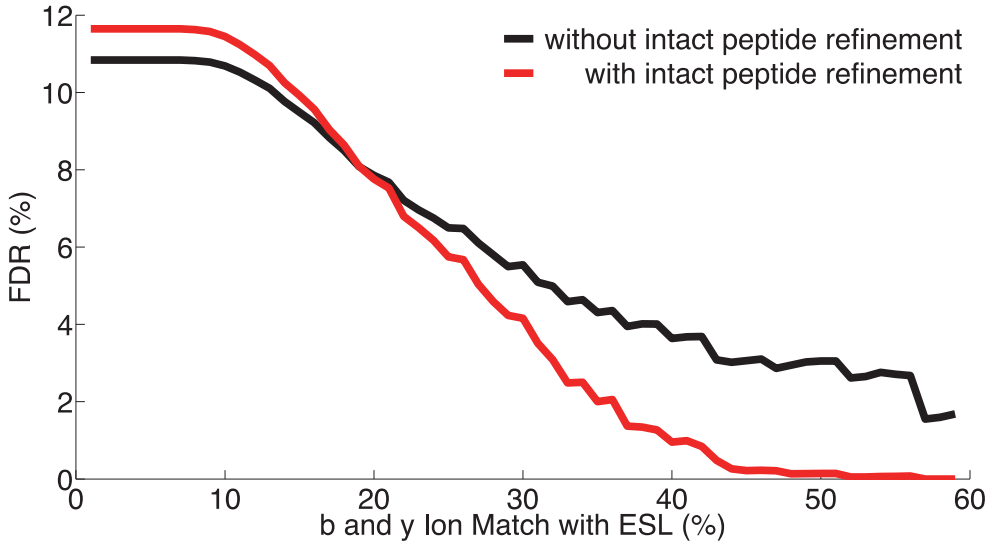


Figure 5-5. Estimation of FDR for glycoproteomics analysis of the LNCaP samples. The FDR was calculated as a function of the percentage of b and y ions in the tandem MS spectra matching the ESL library in the LNCaP cell analysis. The red curve, showing the FDR analysis for the spectral library matching algorithm, shows that an FDR of 1% is achieved by setting this threshold at 40%. The FDR curve in black shows the results of FDR calculation omitting the use of intact peptide ions for refining the results. Comparing the two curves shows that taking the intact peptide ions into account improves the FDR and, subsequently, the specificity of the algorithm.

Applying the aforementioned filters in this study, 344 unique glycosite-containing peptides were matched to 57 N-glycan compositions and 769 unique intact N-glycopeptides were identified from LNCaP cells using the spectral library matching algorithm (Supplementary Table 5-3).

In addition to performing global analysis, using this tool, we can look at the heterogeneity of any glycosite of interest or the glycosylation heterogeneity profile of a sample. Figure 5-6 shows the distribution of different N-glycan compositions in the LNCaP sample, where each bar shows the number of PSMs that matched to a specific N-linked glycan in the sample. According to the glycan profile of LNCaP cells, they contained high-mannose, fucosylated only, sialylated only, fucosylated and sialylated,

and other glycans without fucose and sialic acid in hybrid or complex structures; however, the high-mannose structures were more prevalent. The lower abundance of sialylated glycans could be attributed to instability of sialic acid residues and their loss during sample preparation [65], [81], [113].

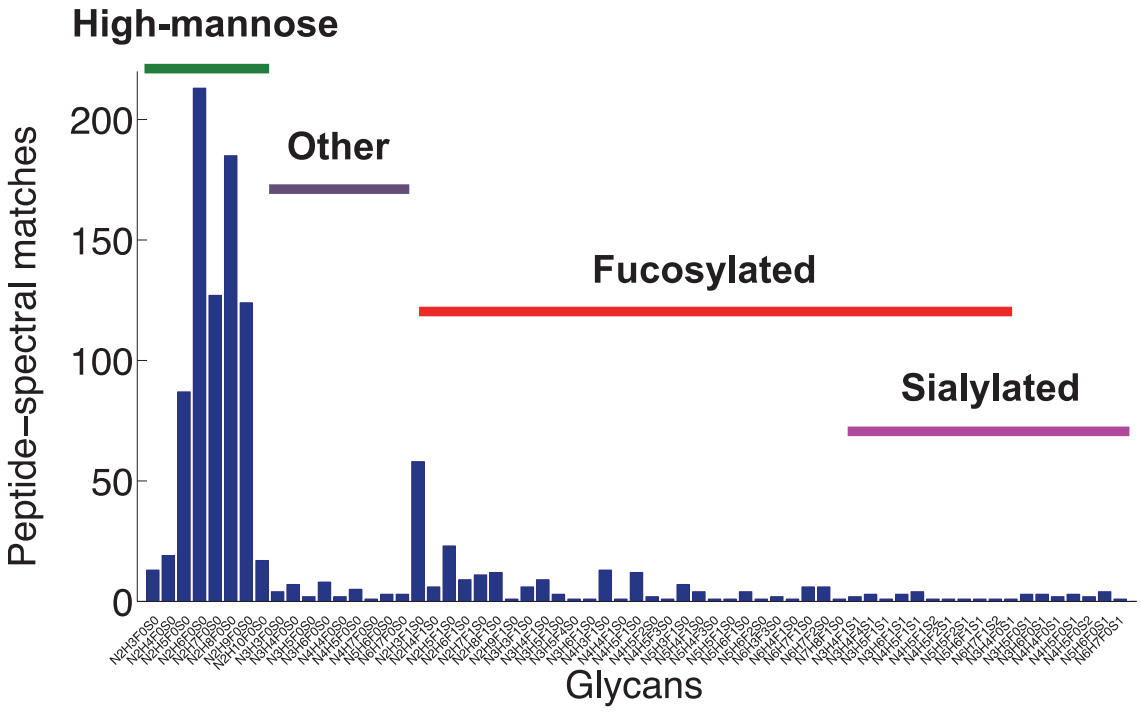


Figure 5-6. Glycan profile of the LNCaP cells.
The number of PSMs pertaining to each glycan is accumulated over all the glycosites. The LNCaP cells contain high-mannose, fucosylated, and sialylated glycans among other N-glycan structures. The high-mannose structures are the most abundant ones identified in the sample, followed by fucosylated glycans. Sialylated glycans are the least abundant structures. Additionally, nonfucosylated, nonsialylated hybrid, and complex N-glycans, marked as others on this figure, were observed in LNCaP cells.

5.4.6 Analysis of unmatched glycan masses in LNCaP samples

Spectral library matching assigns the peptide portion of the intact glycopeptide to the oxonium ion-containing tandem MS spectra, while the decoy strategy ensures the accuracy of the peptide match. The glycan portion is determined by calculating the

glycan mass and matching it to the glycan database. Therefore, if the corresponding glycan structure is missing from the glycan database (Supplementary Table 5-2), the glycan portion of the glycopeptide remains unspecified. This attribute could potentially lead into discovery of novel glycans or glycan modification. In this analysis, we observed examples of LNCaP glycopeptide tandem MS spectra where the peptide portion of the glycopeptides were assigned to glycosite-containing peptides in the ESL, while the calculated glycan masses were missing from the glycan database. As an example, a minimum of 20 spectra in the glycoproteomics analysis of LNCaP cells resulted in a glycan $[M+H]^+$ mass in a 10 ppm window around 1414.512 Da matching to 15 different glycosites; however, the glycan database match did not result in identification of the glycan portion of the intact glycopeptide. In addition, searching the UniCarbKB [133] database for this glycan using the Glycomod tool [134] retrieved no matches. To further analyze the unassigned glycan structure, we re-investigated the tandem MS spectra of LNCaP glycans that had been isolated from the sample and analyzed by HCD LC-MS/MS. Figure 5-7 depicts the tandem MS spectrum of the glycan corresponding to $[M+2H]^{2+}$ mass of 707.76 Da, equivalent to $[M+H]^+$ mass of 1414.512 Da generated through ESI LC-MS/MS analysis of isolated LNCaP glycans on a Q Exactive instrument. The assignment of fragment ions, generated by the Glycoworkbench software [112], to numerous fragmented high-mannose ions suggested that the corresponding unknown glycan was a modified high-mannose structure. In fact, the mass of this glycan is within 10 ppm of a Man6 glycan modified by a moiety with molecular weight of 17.018. This observation, i.e. addition of moiety with a mass of 17.018 Da to a glycan, was observed for Man5, Man7, Man8 and Man9 as well. While an estimated of 700 PSMs matched

unmodified high-mannose structures collectively, about 50 PSMs appeared to match modified high mannose glycans suggesting that around 7% of these structures were modified by an additional 17.018 Da moiety.

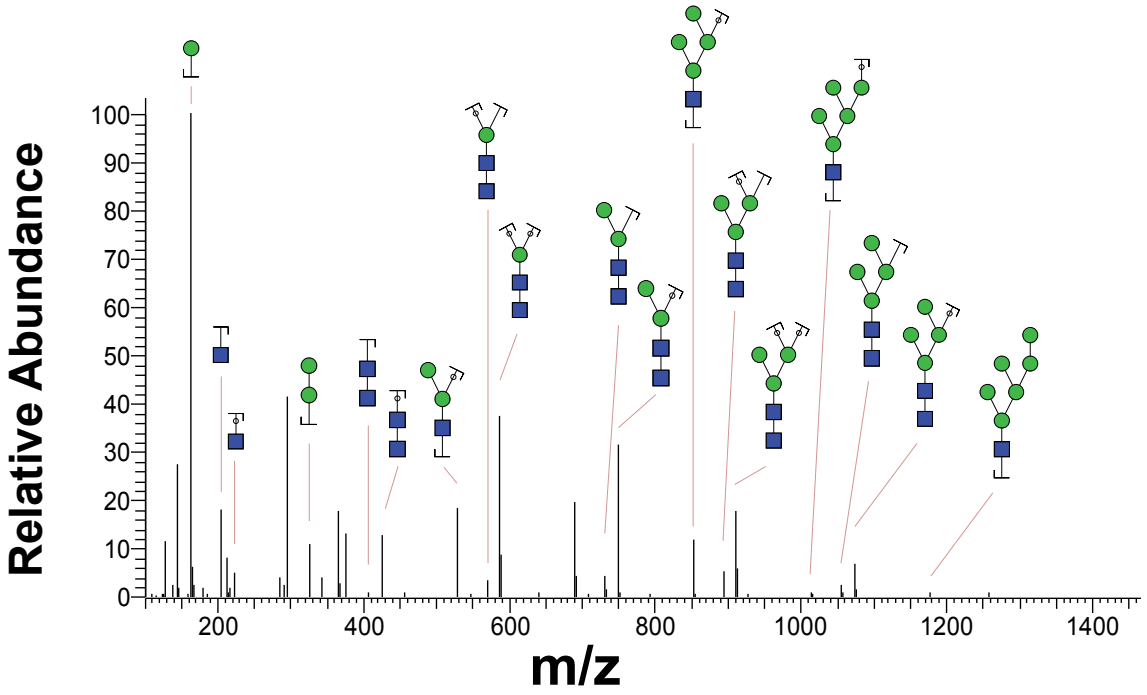


Figure 5-7. Assignment of modified glycans.
Tandem MS spectrum of an uncharacterized glycan peak at [M + H]+ of 1414.512 is assigned to glycan fragment ions. Presence of several oxonium ions in the spectrum of the unknown structure ensures that it belongs to a glycan. Also, assignment of numerous peaks to fragments of Man5 and Man6 glycans suggests that the unknown glycan might be a modified high- mannose structure. Even though the tandem MS spectrum of this glycan is not sufficient to accurately determine the structure, the mass of this structure equals that of Man6 modified by a moiety with a mass of 17.018 Da.

5.5 Discussion and conclusion

Assignment of intact glycopeptides to tandem MS spectra of simple and complex biological samples is a challenging task and several studies have attempted to tackle this challenge. Segu *et al* showed that HCD fragmentation of intact glycopeptides resulted in distinct peptide + GlcNAc ions [85]. Based on this observation, Mayampurath *et al*

developed GlypID 2.0, a tool for assignment of glycopeptide to fragmented glycopeptides [91]. Nwosu *et al* used a combination of accurate precursor mass and the CID-generated fragment ions in tandem MS spectra for assignment of the intact glycopeptides [80]. Singh *et al* took the HCD product ion-triggered ETD of glycopeptide ions and analyzed the ETD fragmented tandem MS spectra using common proteomics tools by introducing the known glycans in the sample as variable modifications [86]. Parker *et al* used proteomics and glycomics techniques to first build databases of possible glycans and glycopeptides in the sample and identified the intact glycopeptides by matching their accurate precursor mass to their concatenated glycopeptide database [129]. He *et al* developed a software tool called GlycoMaster DB for assignment of glycopeptides to HCD/ETD or HCD spectra [99]. These algorithms for glycopeptide assignments assign the tandem spectra based on accurate mass of the precursor ion to potential glycopeptides, which rely on the prior identification of both peptides and glycans in the databases.

Spectral matching of HCD fragmented intact glycopeptides to a tandem mass spectral library of glycosite-containing peptides in a complex sample can be used to reliably identify the peptide portion of these glycopeptides, yielding a false discovery rate in the order of 1%. Knowing the peptide portion, the mass of the glycan portion can be calculated by subtracting the peptide mass from the mass of the precursor ion. Therefore, the glycan structure can be characterized by comparing the glycan mass with the glycan database. Due to various factors, the search for glycan structure might not result in a match. Examples include novelty of the glycan, uncharacterized or sample preparation-induced glycan modifications or errors associated with calculation of the glycan mass. By highlighting these spectra, spectral library matching provides the user with a chance to

further investigate those spectra, where despite reliable peptide identification, the glycan structure remains ambiguous. Using this strategy, our data showed the possibility of a novel modification on glycans with a mass of 17.018 Da. Consideration of this modification could lead to identification of more glycans and intact glycopeptides in biological samples. Similar approaches can be used to identify more uncharacterized glycan modifications.

One of the advantages of spectral library matching is that the accumulation of verified glycopeptide-spectral matches over time facilitates the assignment of newly generated spectra. Furthermore, this concept can be used to group the spectra corresponding to different forms of a peptide to generate a universal fingerprint for that peptide. Comparison of tandem MS spectra to specific universal fingerprints can be facilitated by the many advanced pattern recognition methods developed in the field of signal processing and can result in quick, efficient, and targeted identification of different forms of post-translationally modified peptides in large mass spectrometry datasets.

The human glycoproteome carries the information of not only glycans and proteins, but also glycosylation sites along with the array of glycans present at each site. Our understanding of the glycoproteome is limited, partly due to shortage of the tools available for exploring the glycoproteome. Expanding the glycoproteomics toolkit would provide an opportunity for studying the glycoproteome of specific cells and tissues and revealing changes induced by certain pathologies at molecular level. This knowledge is expected to assist in early detection, accurate diagnosis and improved treatment of human diseases.

Chapter 6. Software-Assisted N- and O-linked Glycoproteomics Analysis Using GPQuest

6.1 Summary

Glycoproteomics, the high-throughput study of protein glycosylation using mass spectrometry, is a rapidly growing field. Recent advances in instrumentation have facilitated the analysis of intact glycopeptides, which is essential for preserving the information of glycosylation microheterogeneity in samples. While massive amount of data is being generated at a rate of ~1 gigabytes/hour on a single instrument, with the software and analytical tools lagging behind, this data is not being fully interrogated. Therefore, there is a great need for analytical tools that can dig the glycoproteomics data. Here we report a software tool for identification of intact N- and O-glycopeptides from simple and complex biological samples, such as recombinant glycoprotein cocktails and cell, serum or tissue samples. GPQuest provides two algorithms named "Precursor Mass Matching" and "Spectral Library Matching" for assignment of `glycopeptides to spectra`, which offer the advantages of lower simulation time and more identifications, respectively. In addition, it uses machine learning to predict the glycosylation type (N- or O-) for each spectrum based on its spectral features with > 90% accuracy. The interface provides the user with the freedom to narrow down the glycopeptide-spectral matches based on oxonium ions, matching b and y ions, and intact peptide ions with partial glycans. Moreover, GPQuest calculates scores for each glycopeptide-spectral match according to the number of matching fragment ions and the intensity coverage of the spectrum by these ions. A decoy strategy was incorporated in GPQuest to validate the results by providing statistics on the false discovery rate. Finally, GPQuest was used for identification of potentially novel O-linked glycosylation sites on bovine fetuin.

6.2 Introduction

Much of the efforts in the glycoproteomics field have focused on N-glycosylation [2], [43], [91], [97]–[99] and little has been done to differentiate between N- and O-glycosylation in the assignment of the tandem MS spectra. Higher-energy collisional dissociation (HCD) alone or in combination with other fragmentation methods has proven to be effective in fragmentation and identification of N-glycoproteins [2], [43], [84], [86], [90]. However, the use of HCD for inspection of O-glycoproteins remains unexplored. The spectra of HCD fragmented O-glycopeptides contain valuable information that require proper analytical tools for interpretation. Assignment of O-glycopeptide structures has been historically more challenging than N-glycopeptides for several reasons. First, lack of a universal enzyme for releasing the O-glycans from their host proteins requires the use of chemical methods such as beta-elimination, which results in suboptimal removal of O-glycans. Second, unlike N-glycosylation sites that are scattered sporadically throughout the protein backbone, O-glycosylation sites tend to be located in close proximities in S/T rich sequences [135]. O-GalNAc type glycoproteins include Serine and Threonine rich short sequences that are heavily O-glycosylated. The proximity of these sites makes it difficult to distinguish the exact structure and site of each O-glycan on the peptide backbone. Last, unlike N-glycosylation where the modification happens on a handful of known motives (NXS, NXT, and less commonly NXC and NXV [32]) O-glycosylation does not target a specific motif, meaning that it can happen on any of the S, T or even Y residues.

Several strategies have been employed for enrichment of glycopeptides. Oxidation of glycans and immobilizing them on solid support has been proven effective for extraction

of glycopeptides [63], [94]. In case of sialylated O-glycoproteins, sialic acids can be selectively oxidized under mild periodate oxidation conditions and O-glycopeptides can be eluted by acid hydrolysis of the sialic acid glycosidic bonds [94]. Clearly, the sialylation information of the sample cannot be retrieved using this strategy. Sialylated glycoproteins are shown to bind to titanium dioxide beads at low-pH. Combining this characteristics with HILIC enrichment, Palmisano *et al* have successfully isolated formerly sialylated glycopeptides from protein mixtures [69]. Lectin weak affinity chromatography (LWAC) is another technique for isolation of O-glycosylated peptides by incorporating different lectins such as wheat germ agglutinin (WGA) [136] and Jacalin [137], [138]. Heterogeneity of O-glycans contributes to the complexity of O-glycoproteome enrichment. To tackle this challenge, a common strategy for isolation of O-linked glycopeptides is engineering the cells to produce simple homogenous O-glycan structures. This is accomplished by targeting the O-glycosylation elongation genes by zinc-finger nuclease, thus creating 'SimpleCell' lines that only express trimmed O-glycans [83], [139]. This technique is valuable in locating the glycosylation sites, however, the O-glycan structures are lost during the process. Once the glycopeptides are collected, they are separated and analyzed by LC-MS/MS and the results are interpreted using glycoproteomics software tools.

Here we report the development of GPQuest, a software tool for analyzing higher-energy collisional dissociation (HCD) fragmented spectra of both N- and O-linked glycopeptides. GPQuest offers two algorithms for matching the tandem mass spectra to intact glycopeptides: precursor mass matching and spectral library matching. This tool supports sample preparation modifications such as stable isotope labeling by amino acids

in cell culture (SILAC) and isobaric tags for relative and absolute quantification (iTRAQ). For ensuring the specificity of the matches, GPQuest uses a decoy strategy to estimate the false discovery rate (FDR) for the analysis [2]. In this strategy, a decoy database is built by concatenating and shuffling the amino acids of the target database and dividing the generated sequence into decoy peptides. The users can choose from three scoring schemes including: fragment ion count score, Morpheus score and intensity score (Table 6-1). A unique feature of GPQuest is its ability to distinguish the tandem mass spectra corresponding to N- and O-linked glycopeptides based on the intensity of oxonium ions. In this strategy, binary classification using logistic regression is used to determine the glycosylation type based on the intensities of 10 oxonium ions resulting from detachment and internal fragmentation of hexose, N-acetylhexosamine, and sialic acid residues from glycans. In this study, we have demonstrated the application of GPQuest for identification of potentially novel O-glycosylation sites on bovine fetuin.

6.3 Methods

6.3.1 GPQuest software development

GPQuest is a software package developed in MATLAB (R2014b) for glycoproteomics analysis of mass spectrometry data [2]. Standalone version of GPQuest, is available for Mac and Windows 64-bit operating systems and can be downloaded at http://www.biomarkercenter.org/GPQuest. The graphical user interface (GUI) of GPQuest is shown in Figure 6-1. The users can choose between two algorithms (precursor mass matching and spectral library matching) for assignment of intact glycopeptides to tandem mass spectra. SILAC and iTRAQ labeling are supported in

GPQuest. Users can extract the intensities of ions (iTRAQ or custom reporter ions) of interest for quantification purposes. The software is designed to accept mzXML2.0 files as input data files and excel or comma-separated value files as glycan and peptide databases. Sample N- and O-linked glycan databases are incorporated in the examples included in the package. In glycoproteomics analysis, one of the first steps is to determine whether a tandem MS spectrum belongs to glycopeptides. This is accomplished by checking the presence of glycan oxonium ions, which are low-mass ions corresponding to mono or disaccharides broken away from the glycan portion of the glycopeptides. The intensities or number of these ions might be affected by fragmentation energy and mass analysis parameters. Therefore, GPQuest gives the user the freedom to determine the filters required to select a spectrum as pertaining to glycopeptides by determining the required minimum number and intensities of the oxonium ions. Furthermore, depending on the type of analysis, the specificity and the glycosylation type, the spectral features vary. Therefore, users can choose what type and charge states of fragment ions are to be considered in their analysis and how many intact peptide ions are expected depending on the length of the glycopeptide of interest. Also, depending on the user's preference three scoring formula for glycopeptide-spectral matches (GPSM) are provided. The fragment ion count scoring takes into account the number of matching peptide fragment ions for each glycopeptide. The Morpheus score focuses on the number of matching fragment and intact peptide ions with a glance at the intensity coverage of the spectra by these ions. The intensity score mainly relies on the intensity coverage of the spectra by the matched ions but also takes the number of these matching ions into account as a secondary

criterion. Last, users can choose the confidence level of analysis by selecting the desired false discovery rate threshold.

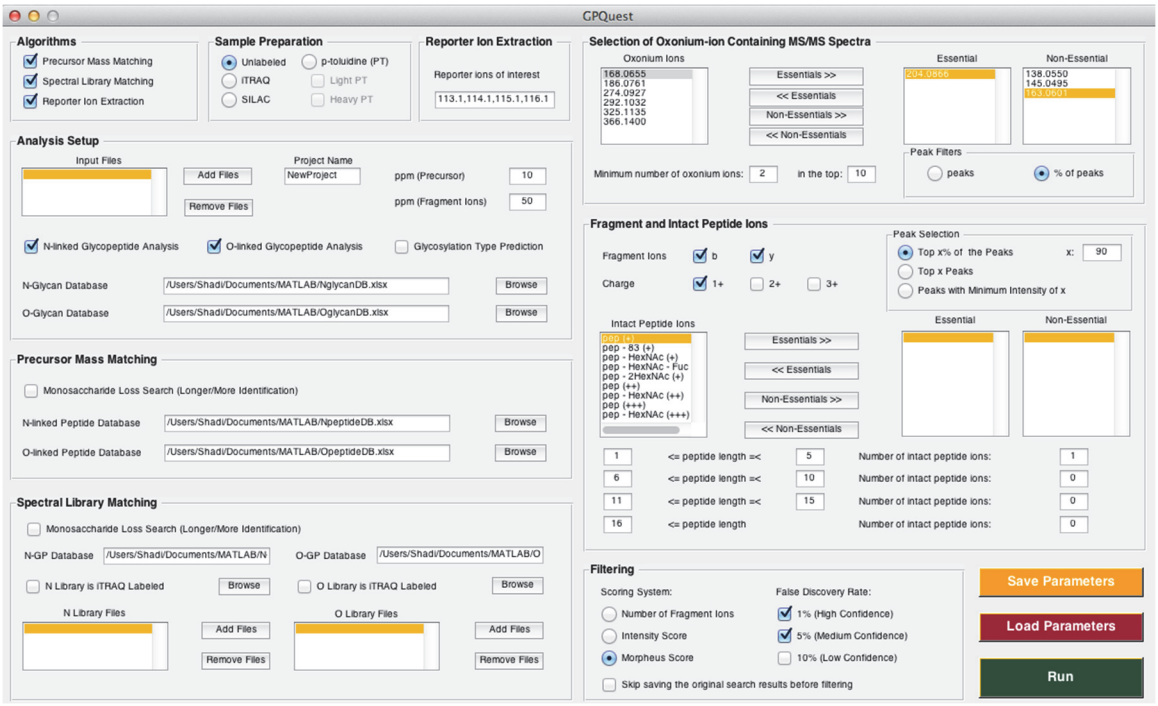


Figure 6-1. Graphical user interface of GPQuest.
The GUI comprises of the following panels: a) algorithms, b) sample preparation, c) reporter ion extraction, d) analysis setup, e) precursor mass matching, f) spectral library matching, g) selection of oxonium ion containing spectra, h) fragment and intact peptide ions and i) filtering.

6.3.1.1 Precursor mass matching and spectral library matching

Precursor mass matching is the assignment of the tandem MS spectra to their corresponding structures by comparing the precursor mass with a glycopeptide database. After the initial matching of the precursor mass, fragment ions are used to score the glycopeptide-spectral matches and refine the search results (Figure 6-2A). In spectral library matching on the other hand, the corresponding peptide to a spectrum can be identified by comparing that spectrum with a spectral library of peptides of interest (Figure 6-2B). Each of these methods offers unique advantages and if used properly can advance high-throughput analysis of mass spectrometry data. For example, precursor

mass matching is the faster of the two algorithms. On the other hand, it is limited to known peptides and glycan structures included in the database. Furthermore, addition of each glycan structure or other post-translational modification (PTM) increases the search space for precursor mass matching analysis. Therefore, this technique is not easily scalable to analysis of multiple PTMs. Precursor mass matching is ideal when a quick solution is desirable and limited number of peptide modifications is under investigation. Spectral library matching is discussed in detail in chapter 5.

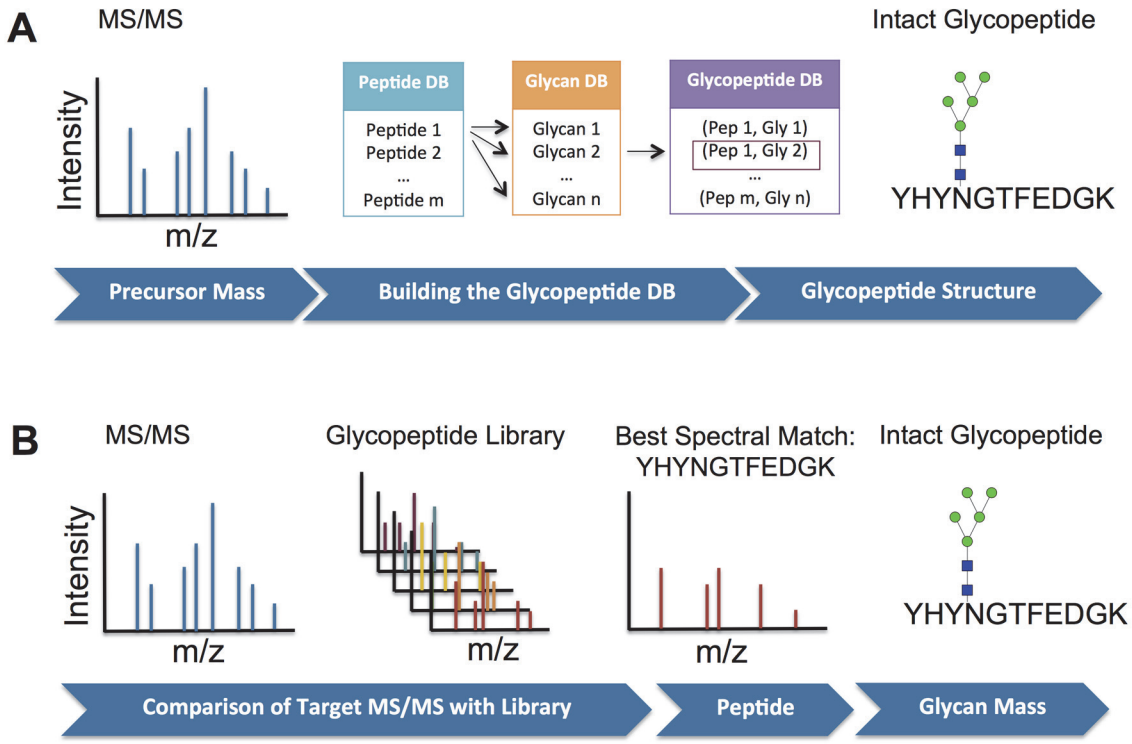


Figure 6-2. Precursor mass matching and spectral library matching for glycopeptide identification.

A) In precursor mass matching a database of potential glycans and glycosite-containing peptides in the sample is used to build a database of theoretical glycopeptides in the sample. Comparing the mass of precursor ion for each tandem MS spectrum generates a list of the most likely corresponding glycopeptides. B) In spectral library matching the tandem MS spectrum is compared with a spectral library of the glycosite-containing peptides in the sample. The closest spectral match pinpoints the most likely peptide corresponding to the spectrum. Knowing the glycopeptide precursor mass and the peptide mass, the glycan mass can be calculated and its structure identified by searching a glycan database.

6.3.1.2 Scoring of glycopeptide-spectral matching

Three scoring schemes are implemented in GPQuest and users can, depending on their analysis and quality of data, choose the scoring system that best fits their needs. These scoring systems are named: 1) Fragment ion count score, intensity score and Morpheus score (Table 6-1). Each scoring method has unique pros and cons as discussed below.

Table 6-1. GPSM score definitions

Score	Definition
Fragment ion count	No. of fragment peptide b and y ions in the spectrum
Morpheus	No. of fragment peptide b and y ions + No. of intact peptide ions + Intensity fraction of matching b, y and intact peptide ions
Intensity	Intensity % of b, y and intact peptide ions + No. of b, y and intact peptide ions

Fragment ion count score is calculated for each GPSM by counting the number of peptide fragment b and y ions that is detected in the tandem MS spectrum. Morpheus score is a modification of the score reported by Wenger *et al* for peptide-spectral matches in high mass resolution proteomics analysis [140]. The original Morpheus score is the sum of the number of fragment b and y ions and the intensity fraction of the tandem MS spectrum covered by the matching fragment ions. For glycoproteomics analysis, we have modified the score in two ways. First, in addition to the fragment b and y ions, this score includes the number of intact peptide ions such as peptide^{+} or peptide-HexNAc^{+}, which are major peaks in HCD fragmented spectra of glycopeptides. Second, the fractional part of the score is the intensity fraction of b, y and intact peptide ions after removing the oxonium ions from the spectra. Excluding the oxonium ion intensities ensures that these ions, which are disproportionally abundant in the spectra of N-glycopeptides, do not

interfere with the scoring of the GPSMs. Intensity score, on the other hand, is designed to emphasize the spectrum intensity coverage by the matching ions as opposed to their number. In particular, this scoring devalues the GPSMs that correspond to randomly matched ions to noisy spectra. Here, the score is calculated by adding the number of b, y and intact peptide ions to the percentage (as opposed to fraction) of the intensity of the tandem MS spectrum covered by the matching ions.

6.3.2 Sample preparation and mass spectrometry analysis

6.3.2.1 Materials and reagents

Sequencing-grade trypsin was purchased from Promega (Madison, WI). C18 desalting cartridges were purchased from Waters (Milford, MA). Glycophorin A (P02724), bovine fetuin (Q58D62 and P12763) and ovalbumin (P01012) were purchased from Sigma Aldrich. All other reagents were purchased from Sigma Aldrich (St. Louis, MO) unless otherwise specified.

6.3.2.2 Glycoprotein sample preparation

The glycoproteins of interest, glycophorin A, ovalbumin and fetuin, were first denatured using 8 M urea in 1 M ammonium bicarbonate buffer with 10 mM tris(2-carboxyethyl)phosphine (TCEP) for 1 hour at room temperature. Denaturing was followed by alkylation using iodoacetamide for 30 min at room temperature in the dark. The alkylated glycoproteins were then digested using high-grade trypsin in a 200 mM ammonium bicarbonate buffer and incubated at 37 °C overnight. The peptides were purified using C18 columns.

6.3.2.3 Liquid chromatography and mass spectrometry analysis

For each glycoprotein, a 0.4 μg aliquot was separated through a C18 column on a Dionex Ultimate 3000 RSLC system and analysed on a Q Exactive mass spectrometer (Thermo Scientific). Data-dependent HCD fragmentation was performed on the 10 most abundant ions using an isolation window of 2 *m/z* and fixed first mass of 100 *m/z*. Unassigned, one, eight and more than eight protonated ions were rejected.

6.3.3 Data analysis

The mass spectrometry raw files were converted to mzXML files using the Trans-Proteomic Pipeline (TPP), with the "centroid all scans" option selected. The precursor mass matching algorithm on GPQuest was used to analyse the mzXML files.

For the fetuin, ovalbumin and glycophorin A samples, the following parameters were used in GPQuest: Mass tolerance was set to 10 and 20 ppm at MS1 and MS2 levels, respectively. N- and O-linked glycosylation analysis were conducted simultaneously. *In silico* tryptic digestion of the amino acid sequences of these three glycoproteins with UniProt IDs of Q58D62, P12763 (fetuin), P02724 (glycophorin A) and P01012 (ovalbumin) was used to generate the peptide database for this analysis. Up to two missed cleavages and maximum two variable oxidations of Methionine were allowed. A tandem MS spectrum was marked as oxonium-ion containing for further glycoproteomics analysis if it contained a minimum of 2 oxonium ions in the 5% highest peaks, with the ion at 204 being mandatory as one of the two. The bottom 10% of the peaks were removed as low-quality peaks and singly charged b and y peptide fragment ions were included for scoring of the GPSMs. For peptides shorter than 11 amino acids, presence of

the intact peptide ion was mandatory. The FDR was set to 1%. For comparison of the three scoring formulas, the presence of intact peptide ion was not mandatory.

6.4 Results

6.4.1 Classification of tandem mass spectra based on glycosylation type

The tandem MS spectra of glycopeptides carry valuable information about the structure that they represent. During HCD fragmentation of glycopeptides, various groups of ions are generated. A key signature of glycopeptide spectra is the presence of oxonium ions that correspond to detached mono- or disaccharides cleaved off the glycans. These oxonium ions not only help distinguish the glycosylated from non-glycosylated peptides, but also could shed light on the structure of their glycan of origin. We have observed that in particular, the oxonium ions released from N- and O-linked glycopeptides were quite different in their intensities and distribution. Figure 6-3 shows the tandem MS spectra of two glycosylated peptides. The top panel shows the spectrum of the peptide 'VTCTLFQTQPVIPQPQPDGAEAEAPSAVPDAAGPTPSAAGPPVASVVVGPSVVA VPLPLHR' which is part of the bovine fetuin sequence and has known O-glycosylation sites. The spectrum shows this peptide modified by an O-linked glycan structure with N3H3F0S1 composition, where N, H, F and S represent the number of HexNAc, Hexose, Fucose and Neu5Ac residues, respectively. The bottom panel depicts the N-glycopeptide 'LCPDCPLLAPLNDSR' that has been modified by N-linked N4H5F0S1. This peptide too belongs to bovine fetuin and has a known N-glycosylation site. Both glycans are sialylated non-fucosylated forms, consist of 7 and 10 monosaccharides respectively, and are of comparable size. The figure also shows the mass/charge range spanning the

oxonium ions and their intensities. Seven ions are annotated on this figure including 138, 168, 186 and 204 (HexNAc ions), 274 and 292 (Neu5Ac ions) and 366 (Hex-HexNAc ion). There are a few notable distinctions between the two spectra. For example, the ratios of the HexNAc ions intensities are significantly different between spectra of N- and O-linked glycopeptides. The Neu5Ac ions have a higher share of the total intensity in the O-linked spectrum compared to the N-linked spectrum, while the opposite is true for the Hex-HexNAc ions.

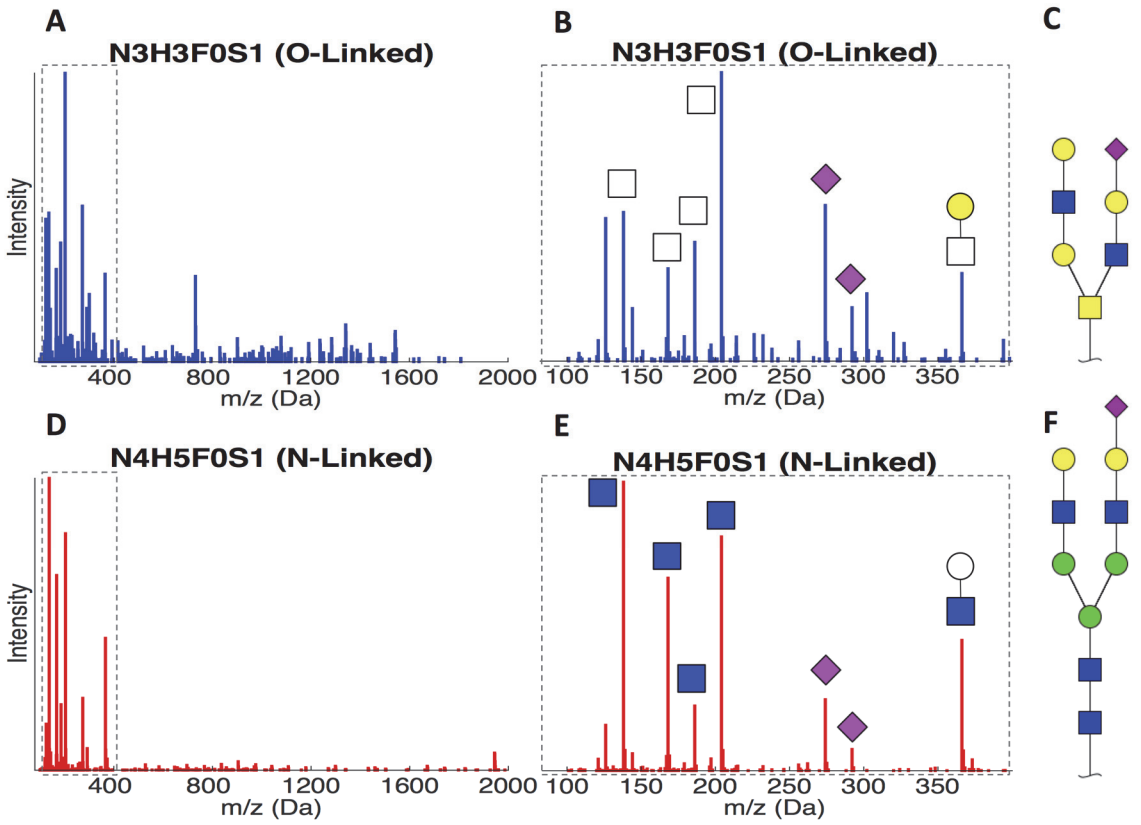


Figure 6-3. Intensity of glycan oxonium ions differs between HCD fragmented O- and N-linked glycopeptides.

The tandem MS spectrum of HCD fragmented glycopeptides with (A) O-linked and (D) N-linked glycans are shown. Focusing on the lower mass range that covers the oxonium ion range for (B) O-linked and (E) N-linked glycans shows that these two glycosylation types create very different patterns of fragment ions in this range, even though the corresponding glycan structures N3H3F0S1 (C) and N4H5F0S1 (F) are of comparable size. There are several differences in the intensities of the oxonium ion between these two glycopeptides. Among these differences are the

intensity ratios of HexNAc related oxonium ions at 138, 168, 186 and 204, the intensities of Neu5Ac related oxonium ions at 274 and 292 and the intensity of HexNAc-Hex oxonium ion at 366.

In order to systematically study the difference between spectral features of N- and O-linked glycopeptides, a training dataset was generated of the HCD fragmented spectra of different glycopeptides. This dataset was built by first analyzing the glycoproteomic data of ovalbumin, glycophorin A and fetuin using GPQuest. Ovalbumin was selected for this experiment because it is an N-glycoprotein, with no reported O-glycosylation sites and therefore provided a number of N-linked glycopeptide spectra for the training dataset. Glycophorin A, on the other hand, is heavily O-glycosylated and therefore provided many O-linked glycopeptide spectra. The fetuin glycoprotein, purchased from Sigma Aldrich, consists of two homologues of fetuin-A and fetuin-B with UniProt IDs of P12763 and Q58D62, both containing reported N- and O-glycosylation sites. Ideally, the training dataset should be diversified to cover a wide variety of possible cases. The size and diversity balance of the O-linked and N-linked entries in the database were improved by adding 79 manually verified O-linked glycopeptide spectra from our study of gp120 glycosylation [90] to this dataset. Several additional filters were applied to GPQuest output to ensure that the spectra selected for the training dataset were of high quality with confident glycopeptide assignments. First of all, only GPSMs including peptides with known glycosylation sites were selected. Also, GPSMs with intensity coverage below 10% and total spectrum intensity below $5x10^5$ were excluded from the dataset. For spectra of O-linked glycopeptides, only GPSMs assigned to peptides lacking any N-glycosylation sites were considered. Concatenation of the data from ovalbumin, glycophorin A, fetuin and gp120 resulted in a training dataset containing 872 and 527

HCD fragmented spectra of N- and O-linked glycopeptide, respectively. FDR was set to 1% for this analysis.

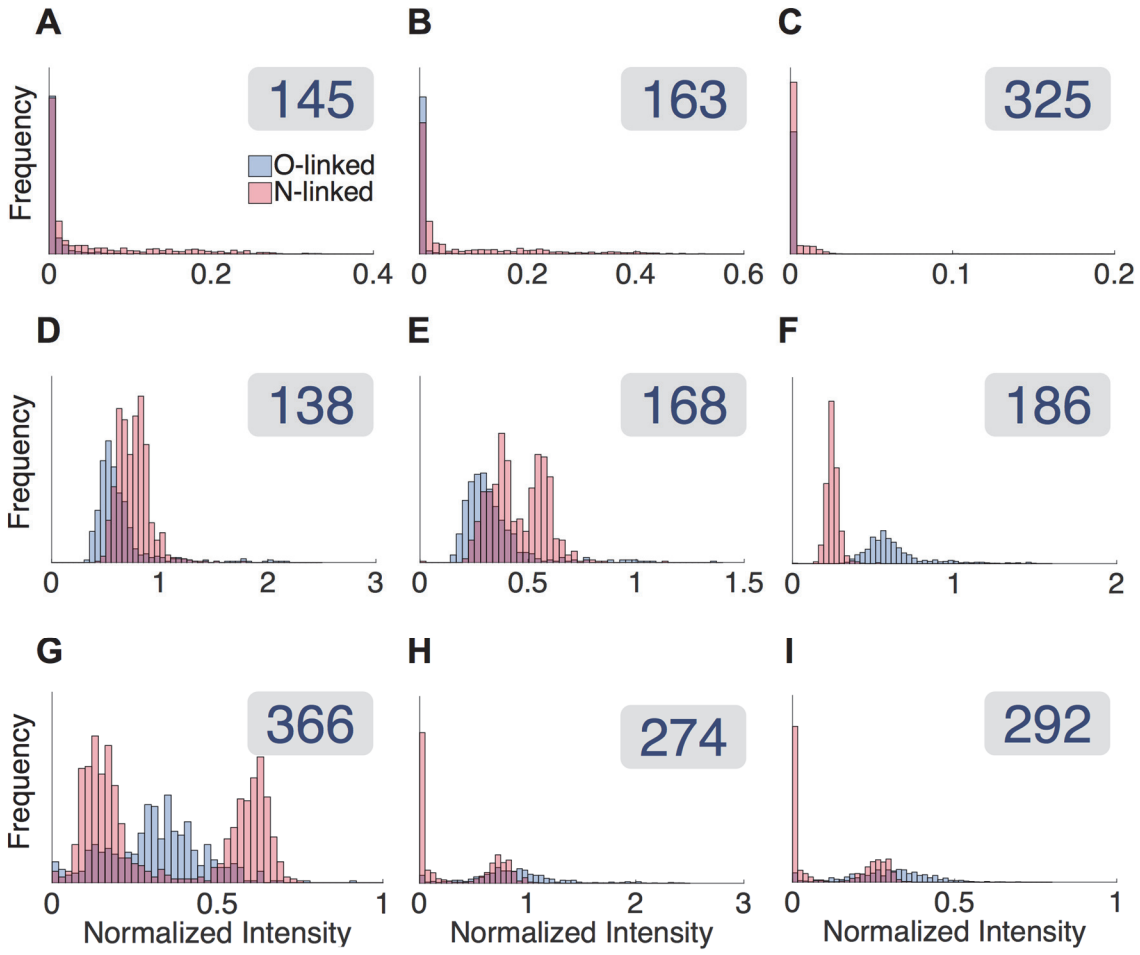


Figure 6-4. Spectral differences between O- and N-linked glycopeptides in the oxonium ion region can be used to predict the glycosylation type for each glycopeptide spectrum.
The intensities of nine different oxonium ions normalized by the intensity of HexNAc ion at 204 are depicted. The plotted ions are as follows: (A) Hex-related ion at 145, (B) Hex ion at 163, (C) Hex-Hex ion at 325, (D) HexNAc-related ion at 138, (E) HexNAc-related ion at 168, (F) HexNAc-related ion at 186, (G) Hex-HexNAc ion at 366, (H) Neu5Ac-related ion at 274 and (I) Neu5Ac ion at 292. Clearly, there are differences between the two types of glycosylation in terms of their oxonium ion intensities. These differences are more pronounced for some of these ions. In particular, normalized intensity of 186 is significantly higher in O-linked spectra. The other two HexNAc related ions are however higher in the N-linked spectra. Furthermore, sialic acid residues have a greater share in the oxonium ion intensities of O-linked spectra while Hex-related oxonium ions are generally higher in the N-linked spectra. These differences, fed into a machine learning algorithm, can help predict the glycosylation type and classify the spectra.

Figure 6-4 illustrates the intensities of oxonium ions in N- and O-linked glycopeptide tandem MS spectra side by side. The intensities were extracted for nine oxonium ions and normalized by the intensity of the tenth ion at m/z 204, which is the intact HexNAc oxonium ion and observed in almost all the spectra that belong to glycopeptides. The distribution of the oxonium ions are plotted for N-linked (in red) and O-linked (in blue) spectra. As shown, there are significant differences between the two types of glycosylation: 1) The Hex related oxonium ions (145, 163 and 325) are usually more abundant in spectra of N-linked glycopeptides. 2) The sialic acid related oxonium ions (274 and 292) are more abundant in spectra of O-linked glycopeptides. 3) There is a clear difference in the normalized intensities of HexNAc related oxonium ions between N- and O- linked glycopeptides. Particularly, the intensity of 186 normalized by 204 is significantly higher in spectra of O-linked glycopeptides compared to those of N-linked ones.

Spectral features can be used to distinguish spectra of N- and O-linked glycopeptides using machine learning. Here, we employed multinomial logistic regression classification, and trained and cross-validated this classifier on our curated spectral dataset of N- and O-linked glycopeptides. The intensities of the nine aforementioned oxonium ions in Figure 6-4 normalized by the intensity of 204 were chosen as the inputs to the classification algorithm. The use of the normalized intensities makes this approach more robust to change of platforms, as different manufacturers and instruments use distinctive units for ion intensities. Classification on this dataset of 1399 spectra from gp120, glycophorin A, ovalbumin and fetuin, yielded an accuracy of 98.8%, where the true ‘N-linked rate’ was 99.7% and the true ‘O-linked rate’ was 97.3%. The classifier was additionally validated

on a second dataset of fetuin, glycophorin A and ovalbumin spectra with 94.4% accuracy. One of the advantages of using logistic regression is that in addition to classifying the spectrum, it provides the numerical probability of the classification being correct. This attribute is particularly crucial for the spectra that lie within the gray zone between the two classes of N- and O-linked. Looking at the statistics of the model coefficients, we observed that the most crucial feature for classification, as predicted, was the intensity ratio of 186 to 204 with a p-value close to zero. We hypothesize that, since O-linked glycans are richer in GalNAc, as opposed to N-linked glycans that are richer in GlcNAc, the change in the ratio of these HexNAc related spectra could point to the dissimilar ways that GalNAc and GlcNAc are fragmented. Halim *et al* showed that the intensities of the oxonium ions in glycopeptide tandem MS spectra help identify the glycan saccharide identities and in fact they demonstrated that the ratio of 186 to 204 oxonium ions is higher for GalNAc containing glycans compared to their GlcNAc containing counterparts [141], which is in line with our observation.

The ability to predict the glycosylation type of a peptide based on its spectral features is beneficial in several ways. First of all, by breaking the search space into subspaces of N- and O-glycopeptides, the analysis could be performed almost twice as fast, which is a significant improvement in the execution time, especially for large datasets. Moreover, it can help distinguish the glycosylation type when the GPSM includes a glycan composition that belongs to both N- and O- glycan databases. For instance, the glycan portion of the glycopeptide corresponding to the spectrum in Figure 6-5 has the composition N3H3F0S0. As reported on the consortium for functional glycomics portal, this composition could potentially represent an N-linked or O-linked glycan (Figure 6-5).

The glycosylation classification model predicts this spectrum to match an O-linked glycopeptide with high probability (>99%). The peptide portion of this GPSM is in fact an O-glycosylated peptide (AHEVSEISVRTVYPPEEETGER) that belongs to glycophorin A and has three reported glycosylation sites. Therefore, in this example we know the prediction of the model to be true. This principle can be applied to predict the glycosylation type on peptides containing both N- and O-glycosylation sites.

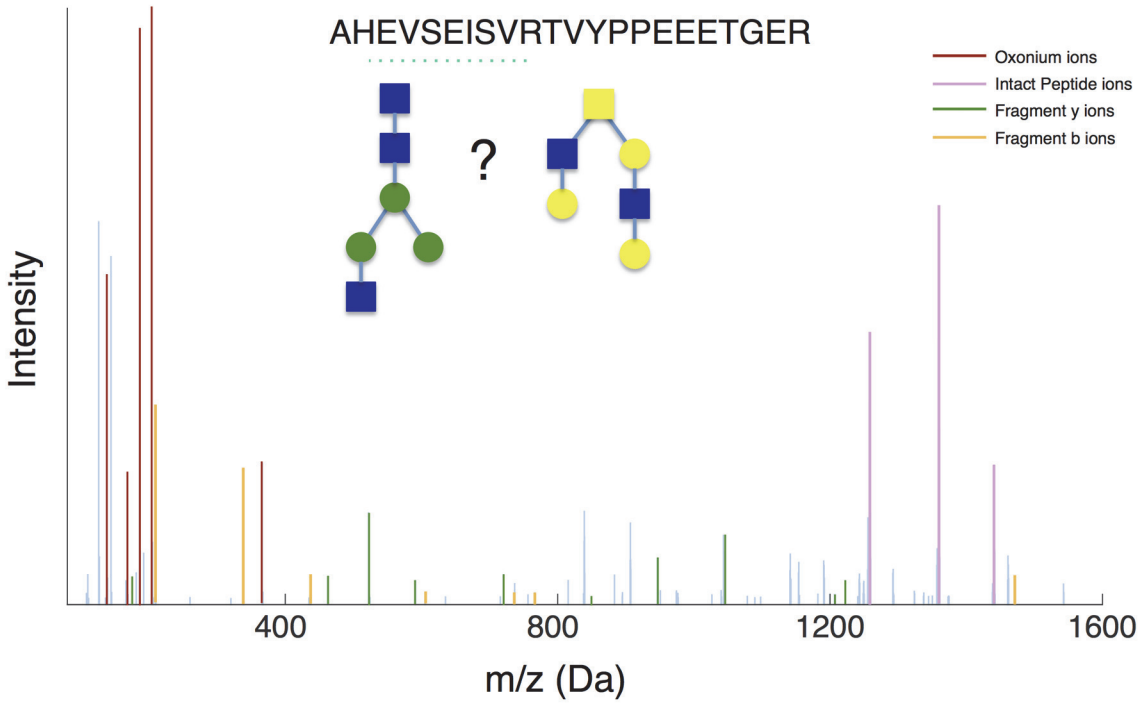


Figure 6-5. Prediction of the glycosylation type.
The tandem MS spectrum corresponding to the following GPSM is shown: AHEVSEISVRTVYPPEEETGER – N3H3F0S0. The glycan composition of this GPSM can represent an N- or O-linked structure. The intensity of the oxonium ions, particularly the high ratio of 186 to 204 implies the glycosylation type to be O-linked, which is in agreement with the peptide sequence, as it lacks N-glycosylation sites and can only be O-glycosylated.

6.4.2 Identification of novel O-glycosylation sites on bovine fetuin

Glycosylation of fetuin-A has been studied extensively using a variety of techniques such as beta-elimination for site-determination, and endoglycosidase digestion, HPLC and LC-MS/MS with ETD or CID fragmentation for structure verification, [79], [87], [137], [142]–[144]. These studies have uncovered three N- (N-99, N-156, N-176) and six O-glycosylation sites (S-271, T-280, S-282, S-296, T334, S-341) for fetuin-A as reported on the UniProt database. The sequence of fetuin-A (P12763) is given in Table 6-2 with the N- and O-glycosites shown in red and blue respectively and the peptides identified using GPQuest marked in green. As shown in Table 6-2, the reported glycosylation sites are covered by the peptides that have been assigned to tandem spectra of fetuin glycopeptides using GPQuest. The aa246-306 sequence starting with VTC and ending with LHR is a large glycopeptide consisting of 61 amino acids and harboring 4 of the reported O-glycosylation sites, i.e. S-271, T-280, S-282 and S-296. The other two reported O-glycosylation sites, T334 and S-341, were covered by aa334-348 or TPIVGQPSIPGGPVR. In addition to these reported sites, GPQuest was able to assign the aa313-333 (HTFSGVASVESSSGEAFHVGK) and aa307-333 (AHYDLRHTFSGVASVESSSGEAFHVGK) sequences to several oxonium ion-containing spectra with high confidence, revealing that these two peptides might in fact be glycosylated. The glycopeptide spectral matches corresponding to these peptides are shown in Table 6-2 along with the list of matching fragment b and y ions for each, which confirms the quality of the match between the glycopeptides and the tandem MS spectra.

Table 6-2. Identification of novel O-glycosites on fetuin-A.
Fetuin-A contains three N- (red) and six O-glycosites (blue) that have been reported in the literature. Identification of glycosylated forms of the aa313-333 and aa307-333 sequences in the

HCD fragmented fetuin tryptic digests suggest that this glycoprotein might harbor an O-glycosite on one or more of the following locations: T-314, S-316, S-320, S-323, S-324, and S-325. The GPSMs corresponding to these potentially O-glycosylated sites are listed in the bottom panel.

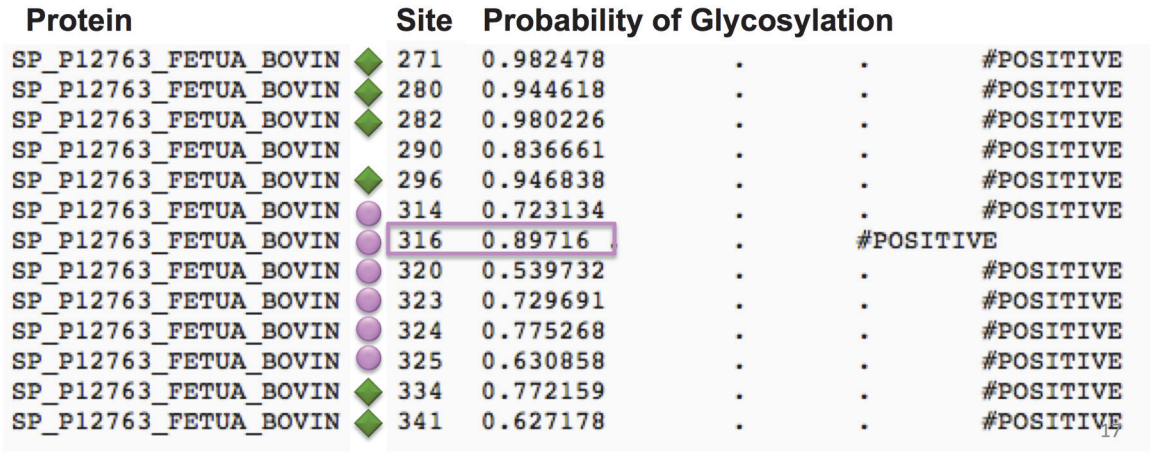
Fetuin-A (P12763) sequence:

MKSFVLLFCLAQLWGCHSIPLDPVAGYKEPACDDPDTEQAALAAVDYINKHLPRGYKHTLNQIDSVKVWP
R[RPTGEVYDIEIDTLETTCHVLDPTPLANCSVR]QQTQHAVEGDCDIHVLKQDGQFSVLFTKCDSSPDSAED
VRK[LCPDCPLLAPLNDSR][VVHAVEVALATFNAESNGSYLQLVEISR]AQFVPLPVSVSVEFAVAATDCIA
KEVVDPTKCNLLAEKQYGFCKGSVIQKALGGEDVR[VTCTLFQTQPVIPQPQPDGAEAEAPSAVPDAAGPT
PSAAGPPVASVVVGPSVVAVPLPLHR][AHYDLRHTFSGVASVESSSGEAFHVGK][TPIVGQPSIPGGPVR]L
CPGRIRYFKI

Peptide	Glycan	Matching b ions	Matching y ions
HTFSGVASVESSSGEAFHVGK	N1H1F0S1	b2/b3/b4/b5/b6/b7/b8/ b9/b10/b11/b14/b16/ b17	y1/y2/y3/y4/y5/y6/y8/y9/ y10/y11/y12/y13/y14/ y15/y17/y18/y19
HTFSGVASVESSSGEAFHVGK	N1H1F0S1	b2/b3/b4/b5/b6/b7/b8/ b9/b10/b11/b12/b13/ b15/b16/b17	y1/y2/y3/y4/y5/y6/y8/y9/ y10/y11/y12/y13/y14/ y15/y16/y18/y19
HTFSGVASVESSSGEAFHVGK	N1H1F0S2	b1/b2/b3/b4/b5/b6/b7/ b8/b9/b10/b12/b15/b16/ b17/b18	y1/y2/y3/y4/y5/y6/y8/y9/ y10/y11/y12/y13/y14/ y15/y17/y18/y19
AHYDLRHTFSGVASVESSSGEAFHVGK	N1H0F0S0		y1/y2/y3/y4/y5/y6/y8/y9/ y10/y11/y12/y13/y14/ y15

Both of the aforementioned sequences harbor S and T amino acids that could be potentially O-glycosylated. These amino acids are highlighted in magenta in Table 6-2. To evaluate the possibility of these peptides being glycosylated, the NetOGlyc tool was applied to the fetuin sequence to predict the possibility of mucin-type O-glycosylation at each site based on its sequence and surface accessibility [145]. The fetuin-A fasta file was submitted to the NetOGlyc 4.0 server for evaluation of the sequence for mucin-type glycosylation sites. The results of O-glycosite prediction of fetuin-A is shown in Figure 6-6. In this figure, the previously reported O-glycosites are marked by a green diamond, while the six potentially glycosites harbored by the aa313-333 sequence are represented by a magenta circle. From NetOGlyc predictions it appears that all of the S and T residues on the latter end of fetuin-A are highly likely

to be glycosylation. In fact, all of the thirteen S and T residues on aa271-357 are predicted to be modified by a mucin-type glycan with a chance of over 50%. Of the thirteen sites shown in Figure 6-6, which only covers aa271-357 of the sequence, six sites have been previously reported and six are within the sequence of aa313-333, which is predicted to be O-glycosylated based on the GPQuest analysis. Among these six sites at positions of 314, 316, 320, 323, 324, and 325, S-316 is the most likely O-glycosite with glycosylation probability of 90%. Targeted ETD fragmentation of the aa313-333 peptide could help identify the exact location of this novel glycosylation site on fetuin-A.



Protein	Site	Probability of Glycosylation			
SP_P12763_FETUA_BOVIN	271	0.982478	.	.	#POSITIVE
SP_P12763_FETUA_BOVIN	280	0.944618	.	.	#POSITIVE
SP_P12763_FETUA_BOVIN	282	0.980226	.	.	#POSITIVE
SP_P12763_FETUA_BOVIN	290	0.836661	.	.	#POSITIVE
SP_P12763_FETUA_BOVIN	296	0.946838	.	.	#POSITIVE
SP_P12763_FETUA_BOVIN	314	0.723134	.	.	#POSITIVE
SP_P12763_FETUA_BOVIN	316	0.89716	.	#POSITIVE	
SP_P12763_FETUA_BOVIN	320	0.539732	.	.	#POSITIVE
SP_P12763_FETUA_BOVIN	323	0.729691	.	.	#POSITIVE
SP_P12763_FETUA_BOVIN	324	0.775268	.	.	#POSITIVE
SP_P12763_FETUA_BOVIN	325	0.630858	.	.	#POSITIVE
SP_P12763_FETUA_BOVIN	334	0.772159	.	.	#POSITIVE
SP_P12763_FETUA_BOVIN	341	0.627178	.	.	#POSITIVE

Figure 6-6. Prediction of O-glycosylation sites on fetuin-A using NetOGlyc.
The known O-glycosylation sites are marked by green diamonds while the potentially O-glycosylated sites, as disclosed by GPQuest analysis, are marked by magenta circles. Based on predictions by NetOGlyc, all six potentially glycosylated sites are likely candidates with probabilities of glycosylation ranging from 54% to 90%. Among the six candidates, S-316 appears to be the most likely glycosylation site with predicted glycosylation chance of 90%.

6.4.3 Scoring of glycopeptide-spectral matches

In order to understand how the scoring strategies affect the output of GPQuest, we analyzed and filtered three datasets of fetuin, glycophorin A and ovalbumin glycoproteins using the fragment ion count, Morpheus and intensity scores, and compared the outputs (Figure 6-7). The number of GPSMs for the three glycoproteins after applying 1%, 5%

and 25% FDR filters based on the three scores is plotted in Figure 6-7A. Based on this comparison, Morpheus and ion count scores appeared to be superior to intensity score in at least two out of three datasets, although Morpheus score slightly outperformed fragment ion count score in the number of identifications. It should be noted that in the ovalbumin datasets at FDRs of 1% and 5%, intensity score performed nearly as well as Morpheus score, suggesting that this scoring might provide an advantage in specific cases or datasets. We further compared the GPSMs assigned in the ovalbumin dataset at 1% using the three scoring systems, the results of which are shown in a Venn diagram in Figure 6-7B. Even though the overlap between GPSMs retuned by filtering based on intensity and Morpheus scores surpassed 80%, there remained to be 38 GPSMs that were only identified when the intensity score was applied. Upon further investigation of spectra corresponding to these 38 GPSMs, we observed that a common feature of these spectra is their sparsity of mass spectral peaks. An instance of a sparse spectrum is shown in Figure 6-7C. In fact, these spectra comprise of ions at fewer than average distinct mass/charge ratios, which tend to match well with the b and y fragment ions of peptides in the database, resulting in high intensity coverage. The sparse nature of these spectra could be due to low abundance of the precursor glycopeptide, or poor fragmentation. In either case, the intensity scoring creates an opportunity to identify these ions and increase identification and therefore we suggest it be used complementary to either Morpheus or fragment ion count scores.

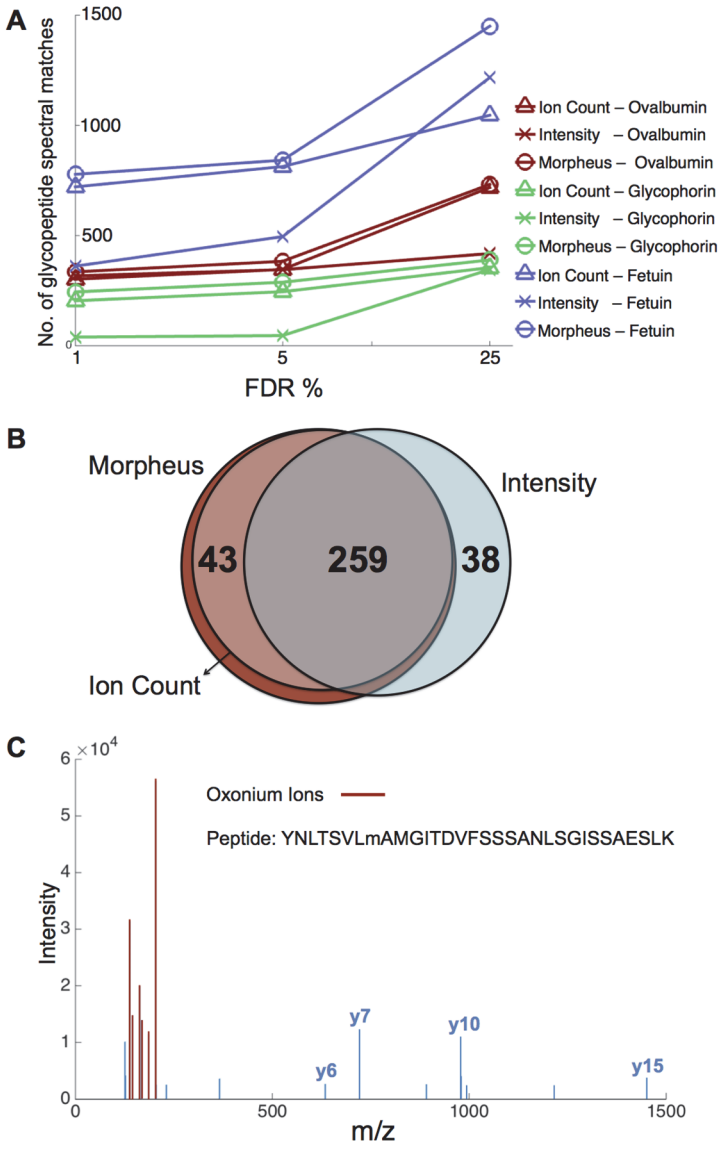


Figure 6-7. Scoring of glycopeptide-spectral matches.

A) The number of GPSMs is shown for filtering of ovalbumin, glycophorin A and fetuin datasets at 1, 5 and 25% FDR using fragment ion count, intensity, and Morpheus scores. In general, Morpheus and intensity score yielded the highest and lowest number of GPSMs, respectively. However, for the ovalbumin dataset, the intensity score was competitive with the other two scores. B) Further comparison of the ovalbumin GPSMs generated by the three scores at 1% FDR showed that there is considerable (>80%) overlap between the GPSMs of intensity and Morpheus scores. However, despite intensity score being inferior to Morpheus score in general, there were 38 GPSMs that could only be identified using the intensity score. C) A representative spectrum of these 38 spectra is shown. The spectrum corresponds to N-glycosylated YNLTSVLMAM[Oxidation]GITDVFSSSANLSGISSAESLK. While only 4 ions were matched to this peptide, these ions covered ~50% of the total spectrum intensity (excluding the oxonium

ions). The intensity score was therefore effective in assignment of such a spectrum that suffers from poor fragmentation and subsequently, small number of matched ions.

6.5 Discussion and conclusion

This study presents a machine learning algorithm for distinguishing the type of glycosylation based on the spectral features of fragmented glycopeptides. This feature are integrated to precursor mass matching and spectral library matching algorithms for software-assisted glycoproteomics analysis of LC-MS/MS data using GPQuest. The algorithms incorporated into GPQuest are designed for analysis of HCD fragmented glycopeptides. HCD has proven to be a powerful technique for N-glycoproteomics analysis [43]; however, its application for assignment of O-glycopeptides is still limited. This is mainly due to the limitations of HCD for determination of the exact glycosite. ETD fragmentation is currently the gold standard for identification of intact O-glycopeptides and the determination of the O-glycosites. The unique feature of ETD in breaking the amide bonds with minimal fragmentation of the post-translational modification makes it an excellent tool for studying PTMs and PTM sites. However, it should be noted that ETD has its own limitations. The quality of ETD fragmentation drastically deteriorates with decreases in charge state and increase in precursor ion mass. In fact, efficiency of ETD decreases for precursor ions with $m/z > 1000$ Da and charge states < 4 [146]. Glycopeptides, due to being modified by large glycans, tend to be in a higher mass range compared to non-modified peptides. In addition, sialic acid can decrease the average charge state of glycopeptides due to its negative charge and since O-glycan are usually heavily sialylated, this characteristics particularly affects the charge state of O-glycopeptides. This could result in non-optimal fragmentation of O-glycopeptides using ETD, which subsequently could compromise the quality of O-

glycopeptide tandem MS spectra and reduce the sensitivity of O-glycopeptide detection. Therefore, the application of HCD for identification of O-glycosylated peptides could be greatly beneficial. In fact, we could take advantage of the superior sensitivity of HCD fragmentation in the glycopeptide mass range for determination of glycopeptide candidates for targeted glycoproteomics approaches using ETD. For example, using HCD, potentially novel O-glycosites were identified in this study that had gone amiss in prior ETD-based glycoproteomics approaches.

Glycan oxonium ions patterns carry much information about the structure, size and composition of their precursor ions. This study focuses on using the intensity of these ions to predict the type of glycosylation in spectra of HCD fragmented glycopeptides. Theoretically, similar principles could be applied to project oxonium ion patterns back to the glycan structure that they represent. Building a spectral library of synthetic glycopeptides with known glycan compositions and structures would be an effective step to study the feasibility of glycan identification based on oxonium ions. This data could potentially lead into computational systems for relating the fragmentation pattern of glycans to their structures. In addition, the current study uses binary classification to predict the glycosylation type. In fact, the classifier divides the space of all possible oxonium ion intensity combinations into two subspaces of N- and O-linked glycosylation. In reality, other classes of glycosylation are possible. The sample preparation method in this study ensures that only glycoproteins are analyzed using LC-MS/MS. However, in general, other glycosylated forms such as proteoglycans, GPI anchors or glycolipids could be analyzed using LC-MS/MS and generate specific oxonium ion patterns that could be used to classify them.

Chapter 7. Future Directions

Glycan imaging is a powerful technique for comparing glycosylation in cells of different origins within the same tissue section. Quantitative mass spectrometry imaging is an emerging area that combines quantitative mass spectrometry strategies with the concept of imaging to construct quantitative images of analyte distributions on tissue sections [147]. Quantitative imaging helps account for the inter-experimental variations and facilitates the comparison of glycan expression among multiple tissue sections. Current mass spectrometry imaging methods rely on label-free quantitation techniques or peak intensities. Therefore, the quantitation accuracy is subject to ionization interference of other co-existing ions. To achieve more reproducible quantification, isotopic or isobaric mass tags could be employed for tissue preparation. Derivatizing the terminal residues on glycans with isotopically labeled reagents provides a means for incorporating isotopic labeling into glycan structure for quantification.

The N-glycan imaging strategy introduced in chapter 4 is restricted to non-sialylated glycans due to the labile nature of sialic acid, which are negatively charged terminal residues on N-linked glycans. There has been evidence to show that sialylation and branching of N-glycans changes in many diseases including cancer [148]–[150]. However, identification and quantification of sialylated glycans is challenging due to the post-source decay of sialic acids in mass spectrometry analysis. Various derivatization techniques have been developed for protection and quantification of sialylated glycans such as permethylation, amidation or esterification. These methods combined with isotopic labeling of glycans can be used for accurate quantification. Permethylation with light and heavy isotopes of methyl groups is an effective method for stabilization of sialic acid. However, the dependence of the subsequent mass shift between the light and heavy

mass peaks on the number of permethylation sites complicates the process of glycan quantification. Alternatively, labeling the reducing end of the glycans with light and heavy isotopes of reagents, resulting in fixed mass shifts between the pairs, can be used for quantification of glycans. However, these techniques fail to derivatize and protect the sialic acids. Shah *et al* have shown that labeling the sialic acid residues with light and heavy isotopes of p-toluidine combines some of the advantage of the two methods by not only stabilizing these residues, but also allowing for quantification of sialylated N-glycans [65]. P-toluidine reacts with the carboxylic groups on sialic acids as well as aspartic acid and glutamic acid on the peptide backbone.

The use of isotopic internal standards in quantitative MALDI imaging has been shown in previous studies [151]–[153]. This technique requires spiking or rather spraying a heavy isotope of the target analyte in a uniform layer over the tissue section. Introduction of internal standards improves the reproducibility and quantitation accuracy of the MALDI mass spectrometry results [154]. *In situ* labeling of proteins and sialylated glycans with light p-toluidine on the tissue and spiking heavy isotope p-toluidine labeled standards helps acquire quantitative mass spectrometry images of not only sialylated glycans but also peptides that contain aspartic or glutamic acids in their sequence (Figure 7-1). Our preliminary study shows that sialylated N-glycans can be effectively labeled with p-toluidine on FFPE tissue and detected using MALDI mass spectrometry. To demonstrate this, we labeled FFPE prostate tissue sections with light p-toluidine in the presence of *N*-(3-Dimethylaminopropyl)-*N*′-ethylcarbodiimide hydrochloride (EDC), released the N-glycans by PNGase F treatment, and profiled the glycans on MALDI-MS. The mass spectrum acquired from the labeled tissue along with that of a non-labeled

control section is illustrated in Figure 7-2. As shown in this figure, labeling of sialylated glycans with p-toluidine protects sialic acid residues from decay, and results in detection and identification of these glycans. In addition, p-toluidine neutralizes the negative charge of sialic acid residues, thus significantly improving the ionization efficiency of sialylated glycans [65]. This approach allows for accurate quantitative imaging of sialylated glycans and proteins without the need for synthetic heavy standards, which are expensive and for most glycans not commercially available.

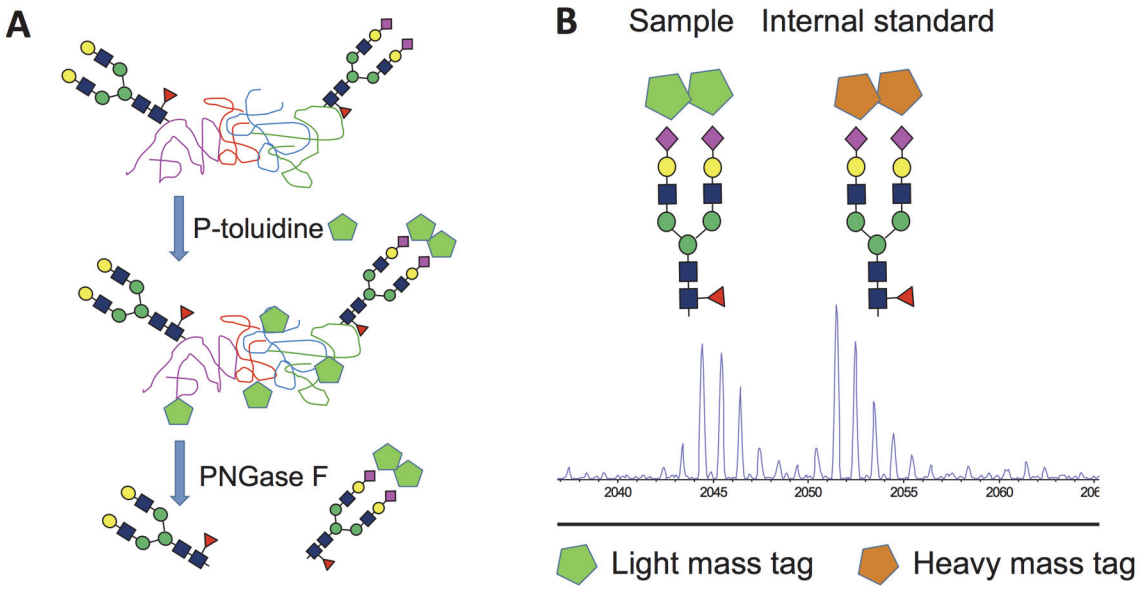


Figure 7-1. Isotope labeling with p-toluidine for glycan quantification.
A) Labeling of glycoproteins with p-toluidine not only protects sialic acid from post-source decay but also allows for quantification of glycans and potentially peptides. B) Glycans or glycoproteins labeled with heavy mass tags are spiked into the sample as internal standards and provide a point of reference for relative or absolute quantification.

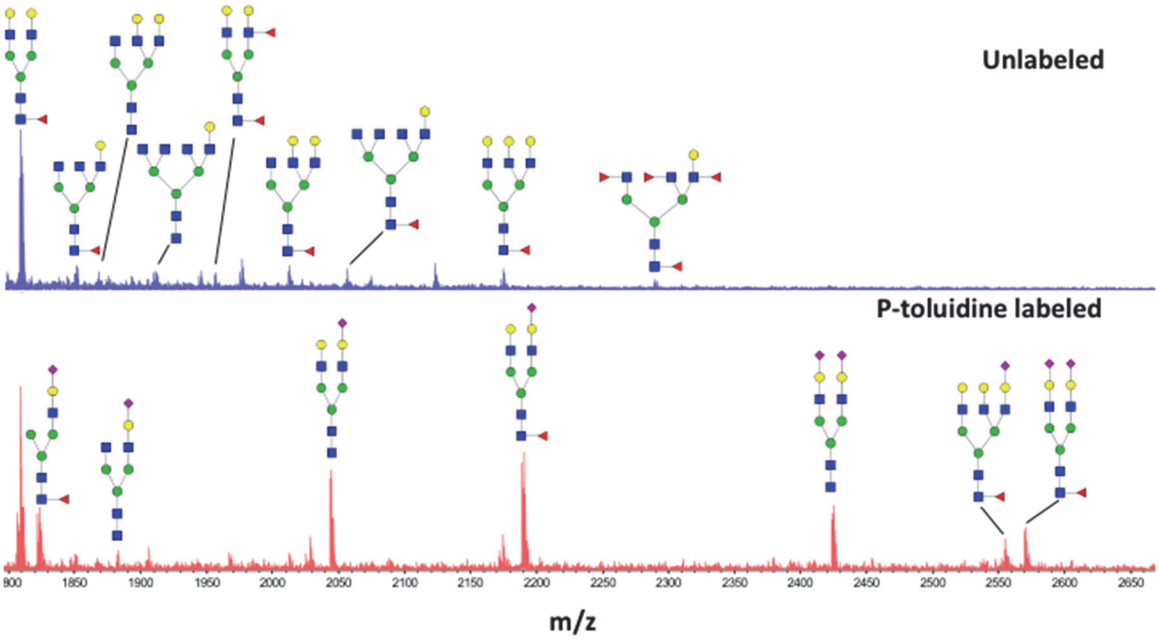


Figure 7-2. In situ labeling of prostate tissue sections with p-toluidine improves detection of sialylated glycans.
P-toluidine improves the detection of sialylated glycans using MALDI-MS by protecting the sialic acid residues from decay during sample preparation or ionization, and neutralizing the negative charge of these analytes, which diminishes the ionization efficiency of glycans in positive mode.

A prospective application of the technologies introduced in this thesis is comprehensive profiling of glycosylation in clinical tissue samples by combining quantitative MALDI imaging of glycans and proteins with site-specific glycoproteomics analysis of the tissue sections. The proposed workflow is presented in Figure 7-3. First, intact glycoproteomics analysis are performed on the tissue sample by in situ tryptic digestion of the proteins followed by LC-MS/MS analysis of the collected peptides and GPQuest. This analysis would identify the intact glycopeptides in each tissue sample, which works as a guide for targeted imaging of the peptides and glycans of interest in the sample. To perform imaging of the glycans, or the peptides, an adjacent tissue section would be labeled with light isotope of p-toluidine, while an external sample, such as

serum will be labeled with heavy isotopes of p-toluidine to be used as internal standards. For this purpose, glycoprotein immobilization for glycan extraction (GIG) [26] and solid phase extraction of glycosite containing peptides (SPEG) [63] could be used for labeling and extraction of serum glycans and peptides, respectively. By applying the isotope-labeled N-glycans or peptides from the serum over the target tissue section, as internal standards, one can achieve relative quantification between sialylated N-glycans or peptides among multiple tissue sections.

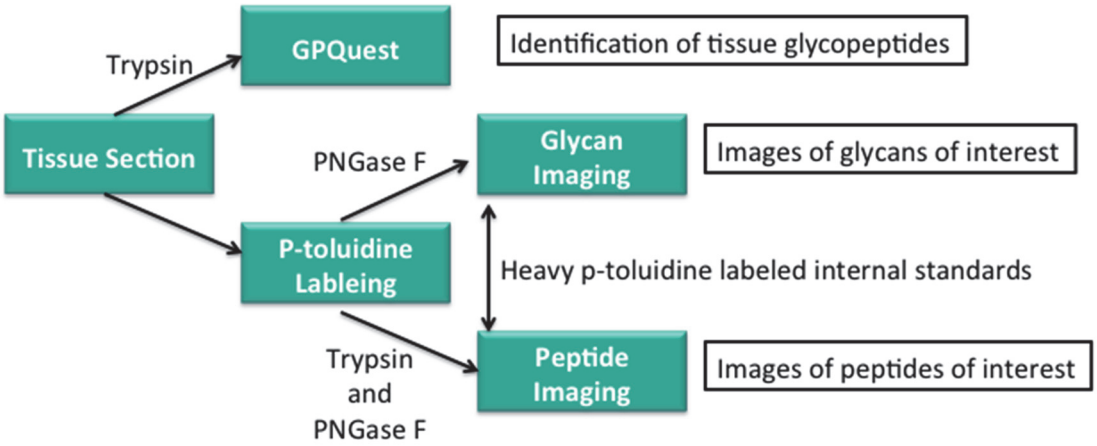


Figure 7-3. Schematic workflow for quantitative imaging of glycans and peptides for comprehensive analysis of glycosylation in tissue sections.
The intact glycopeptides of FFPE tissue sections can be identified by *in situ* tryptic digestion of proteins and glycoproteins directly on tissue and analyzing the extracts with LC-MS/MS and GPQuest. Adjacent sections can be labeled with p-toluidine for quantitative imaging of N-glycans and peptides. Integration of these results could be used to paint a more detailed picture of glycosylation on clinical FFPE tissue samples.

It should be noted that introducing a complex sample as internal standard into the tissue sections could exacerbate the ion suppression effect, thus compromising the signal quality. An alternative solution to overcome this challenge would be to use single glycoprotein or glycopeptides such as sialylglycopeptide to extract isotopically labeled N-glycans or peptides as internal standards. Clearly, this alternative method would limit

accurate quantification to targeted N-glycans and peptides that are present in the standard glycopeptide or glycoprotein mixture. In addition, mass spectral peaks of glycopeptides might be shadowed by the more abundant peptides. Therefore, MALDI might fail to detect intact glycopeptides amongst all the analytes in the sample. In this case, LC-MS/MS analysis is an alternative way to achieve quantitative information of the intact glycopeptides in the sample, though this would clearly compromise the spatial information of the tissue section.

GPQuest has been applied to analysis of numerous samples such as recombinant glycoproteins and cancer cells and have provided invaluable information about glycosylation in these samples [2], [43], [47], [90]. Like any other software tool, the development and evolution of GPQuest is a process and optimization of some of its aspects and functions would improve the overall performance of this tool.

GPQuest focuses on assignment of b and y ion fragments and intact peptide ions with partial fragments. After assignment of a tandem MS spectrum to a glycopeptide and despite a high match score, several peaks remain unassigned. It is likely that these peaks correspond to b and y ions with partial glycans or have resulted from loss of water or ammonia. Inclusion of these modification and glycopeptide fragment ions with multiple fragmentations would potentially improve the identification and scoring of glycopeptide-spectral matches. In addition, assignment of these singular mass peaks might reveal more details about the structure of the glycopeptide such as information about the site of glycosylation.

GPQuest has been developed for assignment of HCD tandem mass spectra of glycopeptides and thus benefits from superior sensitivity of HCD to ETD in the

glycopeptide mass range [146]. However, due to fragmentation of glycosidic bonds in HCD, the precise glycosylation site within the peptide is lost. ETD remains the most effective method for identification of the glycosylation site. HCD fragmentation of glycopeptides combined with GPQuest could be applied for narrowing down the pool of glycopeptides for targeted mass spectrometry analysis of glycopeptides of interest using ETD. Taking advantage of the superior sensitivity of targeted proteomics, this strategy is likely to result in characterization of more O-glycosylation sites.

Glycoproteomics analysis is a computationally demanding task that involves handling of gigabytes of data. Cloud computing is a powerful technology, which is now accessible through resources such as Google cloud or Amazon web services. Our preliminary assessment of parallel processing and cloud computing for glycoproteomics showed that using parallel processing on 4 cores on a local machine or cloud computing on 12 cores on Amazon Elastic Compute Cloud (EC2) could expedite the simulations by ~4 and ~20 times, respectively (Table 7-1). Based on this initial evaluation, integration of cloud computing in GPQuest would potentially improve the execution time and throughput of the process, especially since analysis of large datasets is the most time-consuming step of the glycoproteomics workflow.

Table 7-1. Parallel processing and cloud computing expedites the glycoproteomics simulations using GPQuest.

Number of Cores	Core & Memory	Simulation Time	Comments
1 (Local Machine)	3.7 GHz / 32 GB	~ 90 min	No parallelization
4 (Local Machine)	3.7 GHz / 32 GB	~ 20 min	Requires MATLAB parallel computing toolbox
12 (Amazon EC2)	C3.8xlarge /60 GB	~ 5 min	Requires MATLAB distributed computing server license

Isolation and identification of O-glycosite-containing peptides is an active area of research [83]. Using a sample-specific database of glycosite-containing peptides for precursor ion mass matching of spectral library matching using GPQuest improves the specificity of tandem MS assignment and increase the number of identified glycopeptides [43]. Moreover, the database of potentially O-glycosylated peptides, which comprises of all S/T containing peptides in the complete proteome database, is much larger than a sample-specific database and is computationally expensive. Forthcoming advances in sample preparation for enrichment of O-glycosite containing peptides are expected to improve the specificity, sensitivity and simulation time of glycoproteomics analysis using GPQuest and likely other software tools.

Bibliography

[1] S. T. Eshghi, S. Yang, and X. Wang, "Imaging of N-Linked Glycans from Formalin-Fixed Paraffin-Embedded Tissue Sections Using MALDI Mass Spectrometry," *ACS Chem. Biol.*, vol. 9, no. 9, pp. 2149–2156, 2014.

[2] S. Toghi Eshghi, P. Shah, W. Yang, X. Li, and H. Zhang, "GPQuest: A Spectral Library Matching Algorithm for Site-Specific Assignment of Tandem Mass Spectra to Intact N-glycopeptides," *Anal. Chem.*, vol. 87, pp. 5181–5188, 2015.

[3] F. Crick, "Central dogma of molecular biology.," *Nature*, vol. 227, no. 5258, pp. 561–563, 1970.

[4] A. Varki, A. J. Cummings, J. D. Esko, H. H. Freeze, P. Stanley, C. R. Bertozzi, G. W. Hart, and M. E. Etzler, *Essentials of Glycobiology*, vol. 10, no. 12. Cold Spring Harbor Laboratory Press, 2009.

[5] M. O. N. J. Mann, "Proteomic analysis of post-translational modifications," *Nat. Biotechnol.*, vol. 21, no. 1, pp. 255–261, 2003.

[6] L. Kjeuin and U. Lindahl, "Proteoglycans : Structures," *Annu. Rev. Biochem.*, vol. 60, pp. 443–75, 1991.

[7] G. W. Hart and R. J. Copeland, "Glycomics hits the big time.," *Cell*, vol. 143, no. 5, pp. 672–6, Nov. 2010.

[8] K. W. Moremen, M. Tiemeyer, and A. V Nairn, "Vertebrate protein glycosylation: diversity, synthesis and function.," *Nat. Rev. Mol. Cell Biol.*, vol. 13, no. 7, pp. 448–62, 2012.

[9] "Transforming Glycoscience : A Roadmap for the Future Transforming Glycoscience : A Roadmap for the Future," 2012.

[10] S. Reitsma, D. W. Slaaf, H. Vink, M. A. M. J. Van Zandvoort, and M. G. A. Oude Egbrink, "The endothelial glycocalyx: Composition, functions, and visualization," *Pflugers Arch. Eur. J. Physiol.*, vol. 454, no. 3, pp. 345–359, 2007.

[11] K. Ohtsubo and J. D. Marth, "Glycosylation in cellular mechanisms of health and disease.," *Cell*, vol. 126, no. 5, pp. 855–67, Sep. 2006.

[12] S. R. Stowell, T. Ju, and R. D. Cummings, "Protein Glycosylation in Cancer," *Annu. Rev. Pathol. Mech. Dis.*, vol. 10, no. 1, pp. 473–510, 2015.

[13] S. I. Hakomori, "Aberrant glycosylation in cancer cell membranes as focused on glycolipids: Overview and perspectives," *Cancer Res.*, vol. 45, no. 6, pp. 2405–2414, 1985.

[14] J. W. Dennis, M. Granovsky, and C. E. Warren, "Glycoprotein glycosylation and cancer progression," *Biochim. Biophys. Acta*, vol. 1473, no. 1, pp. 21–34, 1999.

[15] S. Hakomori, "Glycosylation defining cancer malignancy: new wine in an old bottle.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 99, no. 16, pp. 10231–3, Aug. 2002.

[16] M. Ono and S. Hakomori, "Glycosylation defining cancer cell motility and invasiveness.," *Glycoconj. J.*, vol. 20, no. 1, pp. 71–8, 2004.

[17] S. S. Pinho and C. A. Reis, "Glycosylation in cancer: mechanisms and clinical implications," *Nat. Rev. Cancer*, vol. 15, no. 9, pp. 540–555, 2015.

[18] S. M. Grundy, I. J. Benjamin, G. L. Burke, A. Chait, R. H. Eckel, B. V Howard, W. Mitch, S. C. Smith, and J. R. Sowers, "Diabetes and Cardiovascular Disease," *Circulation*, vol. 100, pp. 1134–1146, 1999.

[19] Q. Gong, C. L. Anderson, C. T. January, Z. Zhou, C. L. Anderson, and C. T. Janu-, "Role of glycosylation in cell surface expression and stability of HERG potassium

channels," *Am. J. Physiol. Hear. Circ. Physiol.*, vol. 97201, pp. 77–84, 2002.

[20] N. Fülöp, R. B. Marchase, and J. C. Chatham, "Role of protein O-linked N-acetyl-glucosamine in mediating cell function and survival in the cardiovascular system," *Cardiovasc. Res.*, vol. 73, no. 2, pp. 288–297, 2007.

[21] P. M. Rudd, T. Elliott, P. Cresswell, I. A. Wilson, and R. A. Dwek, "Glycosylation and the immune system," *Science* , vol. 291, no. 5512, pp. 2370–2376, 2001.

[22] D. Chui, G. Sellakumar, R. Green, M. Sutton-Smith, T. McQuistan, K. Marek, H. Morris, a Dell, and J. Marth, "Genetic remodeling of protein glycosylation in vivo induces autoimmune disease.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 98, no. 3, pp. 1142–1147, 2001.

[23] A. Botella-López, F. Burgaya, R. Gavín, M. S. García-Ayllón, E. Gómez-Tortosa, J. Peña-Casanova, J. M. Ureña, J. a Del Río, R. Blesa, E. Soriano, and J. Sáez-Valero, "Reelin expression and glycosylation patterns are altered in Alzheimer's disease.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, no. 14, pp. 5573–5578, 2006.

[24] L. Martin, X. Latypova, and F. Terro, "Post-translational modifications of tau protein: Implications for Alzheimer's disease," *Neurochem. Int.*, vol. 58, no. 4, pp. 458–471, 2011.

[25] S. A. Brooks, "Strategies for analysis of the glycosylation of proteins: Current status and future perspectives," *Mol. Biotechnol.*, vol. 43, no. 1, pp. 76–88, 2009.

[26] S. Yang, Y. Li, P. Shah, and H. Zhang, "Glycomic analysis using glycoprotein immobilization for glycan extraction.," *Anal. Chem.*, vol. 85, no. 11, pp. 5555–61, Jun. 2013.

[27] L. R. Ruhaak, C. Huhn, W. J. Waterreus, A. R. De Boer, C. Neususs, C. H. Hokke,

A. M. Deelder, and M. Wuhrer, “Hydrophilic interaction chromatography-based high-throughput sample preparation method for N-glycan analysis from total human plasma glycoproteins,” *Anal. Chem.*, vol. 80, no. 15, pp. 6119–6126, 2008.

[28] M. Melmer, T. Stangler, A. Premstaller, and W. Lindner, “Comparison of hydrophilic-interaction, reversed-phase and porous graphitic carbon chromatography for glycan analysis,” *J. Chromatogr. A*, vol. 1218, no. 1, pp. 118–123, 2011.

[29] S. Yang, S. T. Eshghi, and H. Chiu, “Glycomic Analysis by Glycoprotein Immobilization for Glycan Extraction and Liquid Chromatography on Microfluidic Chip,” *Anal. Chem.*, vol. 85, no. 21, pp. 10117–10125, 2013.

[30] N. Sharon and H. Lis, *Lectins*, Second. Dordrecht: Springer, 2007.

[31] S. Yang, A. Rubin, S. T. Eshghi, and H. Zhang, “Chemoenzymatic method for glycomics: Isolation, identification, and quantitation,” *Proteomics*, vol. 16, pp. 241–256, 2016.

[32] S. Sun and H. Zhang, “Identification and Validation of Atypical *N* -Glycosylation Sites,” *Anal. Chem.*, vol. 87, no. 24, pp. 11948–11951, 2015.

[33] J. Roth, “Protein N-glycosylation along the Secretory Pathway: Relationship to organelle topography and function, protein quality control, and cell interactions,” *Chem. Rev.*, vol. 102, no. 2, pp. 285–303, 2002.

[34] K. Zarschler, B. Janesch, M. Pabst, F. Altmann, P. Messner, and C. Schaffer, “Protein tyrosine O-glycosylation-A rather unexplored prokaryotic glycosylation system,” *Glycobiology*, vol. 20, no. 6, pp. 787–798, 2010.

[35] M. A. Tarp and H. Clausen, “Mucin-type O-glycosylation and its potential use in

drug and vaccine development," *Biochim. Biophys. Acta*, vol. 1780, no. 3, pp. 546–563, 2008.

[36] L. A. Tabak, "In defense of the oral cavity: structure, biosynthesis, and function of salivary mucins.," *Annu. Rev. Physiol.*, vol. 57, no. 26, pp. 547–564, 1995.

[37] A. Varki, "Glycan-based interactions involving vertebrate sialic-acid-recognizing proteins.," *Nature*, vol. 446, no. 7139, pp. 1023–9, 2007.

[38] G. W. Hart, C. Slawson, G. Ramirez-Correa, and O. Lagerlof, "Cross talk between O-GlcNAcylation and phosphorylation: roles in signaling, transcription, and chronic disease.," *Annu. Rev. Biochem.*, vol. 80, pp. 825–58, 2011.

[39] Y. van Kooyk and G. a Rabinovich, "Protein-glycan interactions in the control of innate and adaptive immune responses.," *Nat. Immunol.*, vol. 9, no. 6, pp. 593–601, 2008.

[40] G. Durand and N. Seta, "Protein glycosylation and diseases: blood and urinary oligosaccharides as markers for diagnosis and therapeutic monitoring.," *Clin. Chem.*, vol. 46, no. 6, pp. 795–805, Jun. 2000.

[41] L. W. Cunningham, R. W. Clouse, and J. D. Ford, "Heterogeneity of the carbohydrate moiety of crystalline ovalbumin," *Biochim. Biophys. Acta*, vol. 78, pp. 379–381, 1963.

[42] G. A. Turner, "N-glycosylation of serum proteins in disease and its investigation using lectins.," *Clin. Chim. Acta.*, vol. 208, no. 3, pp. 149–71, Jun. 1992.

[43] S. Sun, P. Shah, S. T. Eshghi, W. Yang, N. Trikannad, S. Yang, L. Chen, P. Aiyetan, N. Höti, Z. Zhang, D. W. Chan, and H. Zhang, "Comprehensive analysis of protein glycosylation by solid-phase extraction of N-linked glycans and

glycosite-containing peptides," *Nat. Biotechnol.*, vol. 34, no. 1, 2015.

[44] Y. Kaneko, F. Nimmerjahn, and J. V Ravetch, "Anti-inflammatory activity of immunoglobulin G resulting from Fc sialylation.," *Science*, vol. 313, no. 5787, pp. 670–3, Aug. 2006.

[45] G. Vogt, A. Chapgier, K. Yang, N. Chuzhanova, J. Feinberg, C. Fieschi, S. Boisson-Dupuis, A. Alcais, O. Filipe-Santos, J. Bustamante, L. de Beaucoudrey, I. Al-Mohsen, S. Al-Hajjar, A. Al-Ghonaium, P. Adimi, M. Mirsaeidi, S. Khalilzadeh, S. Rosenzweig, O. de la Calle Martin, T. R. Bauer, J. M. Puck, H. D. Ochs, D. Furthner, C. Engelhorn, B. Belohradsky, D. Mansouri, S. M. Holland, R. D. Schreiber, L. Abel, D. N. Cooper, C. Soudais, and J.-L. Casanova, "Gains of glycosylation comprise an unexpectedly large group of pathogenic mutations.," *Nat. Genet.*, vol. 37, no. 7, pp. 692–700, Jul. 2005.

[46] E. Miyoshi, K. Moriwaki, N. Terao, C.-C. Tan, M. Terao, T. Nakagawa, H. Matsumoto, S. Shinzaki, and Y. Kamada, "Fucosylation Is a Promising Target for Cancer Diagnosis and Therapy," *Biomolecules*, vol. 2, no. 1, pp. 34–45, 2012.

[47] P. Shah, X. Wang, W. Yang, S. T. Eshghi, S. Sun, N. Hoti, L. Chen, S. Yang, J. Pasay, A. Rubin, and H. Zhang, "Integrated Proteomic and Glycoproteomic Analyses of Prostate Cancer Cells Reveal Glycoprotein Alteration in Protein Abundance and Glycosylation," *Mol. Cell. Proteomics*, vol. 14, no. 10, pp. 2753–2763, 2015.

[48] E. P. Go, Q. Chang, H. X. Liao, L. L. Sutherland, S. M. Alam, B. F. Haynes, and H. Desaire, "Glycosylation site-specific analysis of clade C HIV-1 envelope proteins," *J. Proteome Res.*, vol. 8, no. 9, pp. 4231–4242, 2009.

[49] E. Ioffe and P. Stanley, “Mice lacking N-acetylglucosaminyltransferase I activity die at mid-gestation, revealing an essential role for complex or hybrid N-linked carbohydrates.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 91, no. 2, pp. 728–32, 1994.

[50] A. Varki, “Sialic acids in human health and disease.,” *Trends Mol. Med.*, vol. 14, no. 8, pp. 351–60, Aug. 2008.

[51] Y.-Y. Zhao, M. Takahashi, J.-G. Gu, E. Miyoshi, A. Matsumoto, S. Kitazume, and N. Taniguchi, “Functional roles of N-glycans in cell signaling and cell adhesion in cancer.,” *Cancer Sci.*, vol. 99, no. 7, pp. 1304–10, Jul. 2008.

[52] H. Schachter and H. H. Freeze, “Glycosylation diseases: quo vadis?,” *Biochim. Biophys. Acta*, vol. 1792, no. 9, pp. 925–30, Sep. 2009.

[53] D. L. Meany and D. W. Chan, “Aberrant glycosylation associated with enzymes as cancer biomarkers.,” *Clin. Proteomics*, vol. 8, no. 1, p. 7, Jan. 2011.

[54] A. Kobata and J. Amano, “Altered glycosylation of proteins produced by malignant cells, and application for the diagnosis and immunotherapyof tumours,” *Immunol. Cell Biol.*, vol. 83, no. 4, pp. 429–439, 2005.

[55] S. Grunewald, “Congenital Disorders of Glycosylation: A Review,” *Pediatr. Res.*, vol. 52, no. 5, pp. 618–624, Oct. 2002.

[56] R. Niehues, M. Hasilik, G. Alton, C. Körner, M. Schiebe-Sukumar, H. G. Koch, K. P. Zimmer, R. Wu, E. Harms, K. Reiter, K. von Figura, H. H. Freeze, H. K. Harms, and T. Marquardt, “Carbohydrate-deficient glycoprotein syndrome type Ib. Phosphomannose isomerase deficiency and mannose therapy.,” *J. Clin. Invest.*, vol. 101, no. 7, pp. 1414–20, 1998.

[57] R. Harold, “Plant lectins : Occurrence , biochemistry , functions and applications,”

Glycobiology, vol. 18, pp. 589–613, 2001.

[58] Y. Li, S. C. Tao, G. S. Bova, A. Y. Liu, D. W. Chan, H. Zhu, and H. Zhang, "Detection and verification of glycosylation patterns of glycoproteins from clinical specimens using lectin microarrays and lectin-based immunosorbent assays," *Anal. Chem.*, vol. 83, no. 22, pp. 8509–8516, 2011.

[59] S. C. Tao, Y. Li, J. Zhou, J. Qian, R. L. Schnaar, Y. Zhang, I. J. Goldstein, H. Zhu, and J. P. Schneck, "Lectin microarrays identify cell-specific and functionally significant cell surface glycan markers," *Glycobiology*, vol. 18, no. 10, pp. 761–769, 2008.

[60] P. Kang, Y. Mechref, I. Klouckova, and M. V. Novotny, "Solid-phase permethylation of glycans for mass spectrometric analysis," *Rapid Commun. Mass Spectrom.*, vol. 19, no. 23, pp. 3421–3428, 2005.

[61] Y. Tian, Y. Zhou, S. Elliott, R. Aebersold, and H. Zhang, "Solid-phase extraction of N-linked glycopeptides.," *Nat. Protoc.*, vol. 2, no. 2, pp. 334–9, Jan. 2007.

[62] S. J. Yang and H. Zhang, "Glycan analysis by reversible reaction to hydrazide beads and mass spectrometry.," *Anal. Chem.*, vol. 84, no. 5, pp. 2232–8, Mar. 2012.

[63] H. Zhang, X.-J. Li, D. B. Martin, and R. Aebersold, "Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry.," *Nat. Biotechnol.*, vol. 21, no. 6, pp. 660–6, Jun. 2003.

[64] Y. Miura, Y. Shinohara, J. I. Furukawa, N. Nagahori, and S. I. Nishimura, "Rapid and simple solid-phase esterification of sialic acid residues for quantitative glycomics by mass spectrometry," *Chem. Eur. J.*, vol. 13, no. 17, pp. 4797–4804,

2007.

[65] P. Shah, S. Yang, S. Sun, P. Aiyetan, K. J. Yarema, and H. Zhang, "Mass spectrometric analysis of sialylated glycans with use of solid-phase labeling of sialic acids.," *Anal. Chem.*, vol. 85, no. 7, pp. 3606–13, Apr. 2013.

[66] L. Royle, C. M. Radcliffe, R. a Dwek, and P. M. Rudd, "Detailed structural analysis of N-glycans released from glycoproteins in SDS-PAGE gel bands using HPLC combined with exoglycosidase array digestions.," *Methods Mol. Biol.*, vol. 347, pp. 125–143, 2006.

[67] A. E. Manzi, K. Norgard-Sumnicht, S. Argade, J. D. Marth, H. van Halbeek, and A. Varki, "Exploring the glycan repertoire of genetically modified mice by isolation and profiling of the major glycan classes and nano-NMR analysis of glycan mixtures.," *Glycobiology*, vol. 10, no. 7, pp. 669–689, 2000.

[68] W. R. Alley and M. V Novotny, "Structural Glycomic Analyses at High Sensitivity: A Decade of Progress," *Annu. Rev. Anal. Chem.*, vol. 6, pp. 237–265, 2013.

[69] G. Palmisano, S. E. Lendal, K. Engholm-Keller, R. Leth-Larsen, B. L. Parker, and M. R. Larsen, "Selective enrichment of sialic acid-containing glycopeptides using titanium dioxide chromatography with analysis by HILIC and mass spectrometry," *Nat Protoc*, vol. 5, no. 12, pp. 1974–1982, 2010.

[70] P. H. Jensen, N. G. Karlsson, D. Kolarich, and N. H. Packer, "Structural analysis of N- and O-glycans released from glycoproteins," *Nat. Protoc.*, vol. 7, no. 7, pp. 1299–1310, 2012.

[71] M. Wuhrer, A. M. Deelder, and C. H. Hokke, "Protein glycosylation analysis by

liquid chromatography-mass spectrometry," *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.*, vol. 825, no. 2, pp. 124–133, 2005.

[72] A. Berthod, S. S. C. Chang, J. P. S. Kullman, and D. W. Armstrong, "Practice and mechanism of HPLC oligosaccharide separation with a cyclodextrin bonded phase," *Talanta*, vol. 47, no. 4, pp. 1001–1012, 1998.

[73] J. Zaia, "Mass Spectrometry and Glycomics," *Omi. A J. Integr. Biol.*, vol. 14, no. 4, pp. 401–418, 2010.

[74] M. Wuhrer, M. I. Catalina, A. M. Deelder, and C. H. Hokke, "Glycoproteomics based on tandem mass spectrometry of glycopeptides.," *J. Chromatogr. B. Analyt. Technol. Biomed. Life Sci.*, vol. 849, no. 1–2, pp. 115–28, Apr. 2007.

[75] S. Sekiya, Y. Wada, and K. Tanaka, "Derivatization for Stabilizing Sialic Acids in MALDI-MS," *Anal. Chem.*, vol. 77, no. 15, pp. 4962–4968, 2005.

[76] S. J. North, P. G. Hitchen, S. M. Haslam, and A. Dell, "Mass spectrometry in the analysis of N-linked and O-linked glycans," *Curr. Opin. Struct. Biol.*, vol. 19, no. 5, pp. 498–506, 2009.

[77] Y. Wada, P. Azadi, C. E. Costello, A. Dell, R. A. Dwek, H. Geyer, R. Geyer, K. Kakehi, N. G. Karlsson, K. Kato, N. Kawasaki, K.-H. Khoo, S. Kim, A. Kondo, E. Lattova, Y. Mechref, E. Miyoshi, K. Nakamura, H. Narimatsu, M. V. Novotny, N. H. Packer, H. Perreault, J. Peter-Katalinic, G. Pohlentz, V. N. Reinhold, P. M. Rudd, A. Suzuki, and N. Taniguchi, "Comparison of the Methods for Profiling Glycoprotein Glycans : HUPO HGPI (Human Proteome Organisation Human Disease Glycomics / Proteome Initiative) Multi-institutional Study," *Med. Phys.*, vol. 17, no. 4, pp. 411–422, 2007.

[78] Y. Wada, M. Tajiri, and S. Yoshida, “Hydrophilic affinity isolation and MALDI multiple-stage tandem mass spectrometry of glycopeptides for glycoproteomics,” *Anal. Chem.*, vol. 76, no. 22, pp. 6560–6565, 2004.

[79] M. Thaysen-Andersen, S. Mysling, and P. Højrup, “Site-specific glycoprofiling of N-linked glycopeptides using MALDI-TOF MS: Strong correlation between signal strength and glycoform quantities,” *Anal. Chem.*, vol. 81, no. 10, pp. 3933–3943, 2009.

[80] C. C. Nwosu, R. R. Seipert, J. S. Strum, S. S. Hua, H. J. An, A. M. Zivkovic, B. J. German, and C. B. Lebrilla, “Simultaneous and extensive site-specific N- and O-glycosylation analysis in protein mixtures,” *J. Proteome Res.*, vol. 10, no. 5, pp. 2612–2624, 2011.

[81] D. J. Harvey, “Matrix-assisted laser desorption/ionization mass spectrometry of carbohydrates.,” *Mass Spectrom. Rev.*, vol. 18, no. 6, pp. 349–450, 1999.

[82] C. S. Ho, C. W. K. Lam, M. H. M. Chan, R. C. K. Cheung, L. K. Law, L. C. W. Lit, K. F. Ng, M. W. M. Suen, and H. L. Tai, “Electrospray ionisation mass spectrometry: principles and clinical applications.,” *Clin. Biochem.*, vol. 24, no. 1, pp. 3–12, 2003.

[83] S. B. Levery, C. Steentoft, A. Halim, Y. Narimatsu, H. Clausen, and S. Y. Vakhrushev, “Advances in mass spectrometry driven O-glycoproteomics,” *Biochim. Biophys. Acta*, vol. 1850, no. 1, pp. 33–42, 2015.

[84] G. Hart-Smith and M. J. Raftery, “Detection and characterization of low abundance glycopeptides via higher-energy C-trap dissociation and orbitrap mass analysis.,” *J. Am. Soc. Mass Spectrom.*, vol. 23, no. 1, pp. 124–40, Jan. 2012.

[85] Z. Segu and Y. Mechref, “Characterizing protein glycosylation sites through higher-energy C-trap dissociation,” *Rapid Commun. Mass Spectrom.*, vol. 24, pp. 1217–1225, 2010.

[86] C. Singh, C. G. Zampronio, A. J. Creese, and H. J. Cooper, “Higher energy collision dissociation (HCD) product ion-triggered electron transfer dissociation (ETD) mass spectrometry for the analysis of N-linked glycoproteins.,” *J. Proteome Res.*, vol. 11, no. 9, pp. 4517–25, Sep. 2012.

[87] M. Windwarder and F. Altmann, “Site-specific analysis of the O-glycosylation of bovine fetuin by electron-transfer dissociation mass spectrometry,” *J. Proteomics*, vol. 108, pp. 258–268, 2014.

[88] K. B. Chandler, P. Pompach, R. Goldman, and N. Edwards, “Exploring site-specific N-glycosylation microheterogeneity of haptoglobin using glycopeptide CID tandem mass spectra and glycan database search.,” *J. Proteome Res.*, vol. 12, no. 8, pp. 3652–66, Aug. 2013.

[89] Y. Mechref, “Use of CID/ETD Mass Spectrometry to Analyze Glycopeptides,” *Curr. Protoc. Protein Sci.*, 2012.

[90] W. Yang, P. Shah, S. Toghi Eshghi, S. Yang, S. Sun, M. Ao, A. Rubin, J. B. Jackson, and H. Zhang, “Glycoform analysis of recombinant and human immunodeficiency virus envelope protein gp120 via higher energy collisional dissociation and spectral-aligning strategy.,” *Anal. Chem.*, vol. 86, no. 14, pp. 6959–67, Jul. 2014.

[91] A. M. Mayampurath, Y. Wu, Z. M. Segu, Y. Mechref, and H. Tang, “Improving confidence in detection and characterization of protein N-glycosylation sites and

microheterogeneity.," *Rapid Commun. Mass Spectrom.*, vol. 25, no. 14, pp. 2007–19, Jul. 2011.

[92] N. E. Scott, B. L. Parker, A. M. Connolly, J. Paulech, A. V. G. Edwards, B. Crossett, L. Falconer, D. Kolarich, S. P. Djordjevic, P. Højrup, N. H. Packer, M. R. Larsen, and S. J. Cordwell, "Simultaneous glycan-peptide characterization using hydrophilic interaction chromatography and parallel fragmentation by CID, higher energy collisional dissociation, and electron transfer dissociation MS applied to the N-linked glycoproteome of Campylobact," *Mol. Cell. Proteomics*, vol. 10, no. 2, pp. M000031–MCP201, Mar. 2011.

[93] H. Kaji, H. Saito, Y. Yamauchi, T. Shinkawa, M. Taoka, J. Hirabayashi, K. Kasai, N. Takahashi, and T. Isobe, "Lectin affinity capture, isotope-coded tagging and mass spectrometry to identify N-linked glycoproteins.," *Nat. Biotechnol.*, vol. 21, no. 6, pp. 667–672, 2003.

[94] J. Nilsson, U. Rüetschi, A. Halim, C. Hesse, E. Carlsohn, G. Brinkmalm, and G. Larson, "Enrichment of glycopeptides for glycan structure and attachment site identification.," *Nat. Methods*, vol. 6, no. 11, pp. 809–811, 2009.

[95] D. Kolarich, P. H. Jensen, F. Altmann, and N. H. Packer, "Determination of site-specific glycan heterogeneity on glycoproteins," *Nat Protoc*, vol. 7, no. 7, pp. 1285–1298, 2012.

[96] M. Bern, Y. J. Kil, and C. Becker, "Byonic: advanced peptide and protein identification software.," *Curr. Protoc. Bioinformatics*, vol. 13, no. 10, pp. 1–17, 2012.

[97] S.-W. Wu, S.-Y. Liang, T.-H. Pu, F.-Y. Chang, and K.-H. Khoo, "Sweet-Heart - an

integrated suite of enabling computational tools for automated MS2/MS3 sequencing and identification of glycopeptides.," *J. Proteomics*, vol. 84, pp. 1–16, 2013.

[98] K. Khatri, G. O. Staples, N. Leymarie, D. R. Leon, L. Turiák, Y. Huang, S. Yip, H. Hu, C. F. Heckendorf, and J. Zaia, "Confident assignment of site-specific glycosylation in complex glycoproteins in a single step.," *J. Proteome Res.*, vol. 13, no. 10, pp. 4347–55, 2014.

[99] L. He, L. Xin, B. Shan, G. A. Lajoie, and B. Ma, "GlycoMaster DB: Software to Assist the Automated Identification of N-Linked Glycopeptides by Tandem Mass Spectrometry," *J. Proteome Res.*, vol. 13, no. 9, pp. 3881–3895, 2014.

[100] H. Hu, K. Khatri, and J. Zaia, "ALGORITHMS AND DESIGN STRATEGIES TOWARDS AUTOMATED GLYCOPROTEOMICS ANALYSIS," *Mass Spectrom. Rev.*, 2016.

[101] H. Debray, D. Decout, G. Strecker, G. Spik, and J. Montreuil, "Specificity of twelve lectins towards oligosaccharides and glycopeptides related to N-glycosylproteins.," *Fed. Eur. Biochem. Soc. J.*, vol. 117, no. 1, pp. 41–55, 1981.

[102] M. Stoeckli, P. Chaurand, D. E. Hallahan, and R. M. Caprioli, "Imaging mass spectrometry: a new technology for the analysis of protein expression in mammalian tissues.," *Nat. Med.*, vol. 7, no. 4, pp. 493–6, Apr. 2001.

[103] E. H. Seeley, S. R. Oppenheimer, D. Mi, P. Chaurand, and R. M. Caprioli, "Enhancement of protein sensitivity for MALDI imaging mass spectrometry after chemical treatment of tissue sections.," *J. Am. Soc. Mass Spectrom.*, vol. 19, no. 8, pp. 1069–77, Aug. 2008.

[104] N. Goto-Inoue, T. Hayasaka, N. Zaima, and M. Setou, “Imaging mass spectrometry for lipidomics.,” *Biochim. Biophys. Acta*, vol. 1811, no. 11, pp. 961–9, Nov. 2011.

[105] D. S. Cornett, S. L. Frappier, and R. M. Caprioli, “MALDI-FTICR imaging mass spectrometry of drugs and metabolites in tissue.,” *Anal. Chem.*, vol. 80, no. 14, pp. 5648–53, Jul. 2008.

[106] M. R. Groseclose, P. P. Massion, P. Chaurand, and R. M. Caprioli, “High-throughput proteomic analysis of formalin-fixed paraffin-embedded tissue microarrays using MALDI imaging mass spectrometry.,” *Proteomics*, vol. 8, no. 18, pp. 3715–24, Sep. 2008.

[107] J. Zaia, “Mass spectrometry and the emerging field of glycomics.,” *Chem. Biol.*, vol. 15, no. 9, pp. 881–92, Sep. 2008.

[108] A. Dell, A. J. Reason, K. H. Khoo, M. Panico, R. A. McDowell, and H. R. Morris, “Mass spectrometry of carbohydrate-containing biopolymers.,” *Methods Enzymol.*, vol. 230, no. 1991, pp. 108–32, Jan. 1994.

[109] J. Zaia, “Mass spectrometry of oligosaccharides.,” *Mass Spectrom. Rev.*, vol. 23, no. 3, pp. 161–227, 2004.

[110] K. L. Chaichana, H. Guerrero-cazares, V. Capilla-gonzalez, G. Zamora-berridi, P. Achanta, O. Gonzalez-perez, G. I. Jallo, J. M. Garcia-verdugo, and A. Qui, “Intra-operatively obtained human tissue : Protocols and techniques for the study of neural stem cells,” *J. Neurosci. Methods*, vol. 180, pp. 116–125, 2009.

[111] T. J. Garrett, M. C. Prieto-Conaway, V. Kovtoun, H. Bui, N. Izgarian, G. Stafford, and R. a. Yost, “Imaging of small molecules in tissue sections with a new

intermediate-pressure MALDI linear ion trap mass spectrometer," *Int. J. Mass Spectrom.*, vol. 260, no. 2–3, pp. 166–176, Feb. 2007.

[112] A. Ceroni, K. Maass, H. Geyer, R. Geyer, A. Dell, and S. M. Haslam, "GlycoWorkbench: a tool for the computer-assisted annotation of mass spectra of glycans," *J. Proteome Res.*, vol. 7, no. 4, pp. 1650–1659, 2008.

[113] S. F. Wheeler, P. Domann, and D. J. Harvey, "Derivatization of sialic acids for stabilization in matrix-assisted laser desorption/ionization mass spectrometry and concomitant differentiation of alpha(2 --> 3)- and alpha(2 --> 6)-isomers.," *Rapid Commun. Mass Spectrom.*, vol. 23, no. 2, pp. 303–312, 2009.

[114] K. A. Stumpo and V. N. Reinhold, "The N-Glycome of Human Plasma research articles," *J. Proteome Res.*, vol. 9, pp. 4823–4830, 2010.

[115] D. Aldredge, H. J. An, N. Tang, K. Waddell, and C. B. Lebrilla, "Annotation of a serum N-glycan library for rapid identification of structures.," *J. Proteome Res.*, vol. 11, no. 3, pp. 1958–68, Mar. 2012.

[116] Y. Hu, S. Zhou, S. I. Khalil, C. L. Renteria, and Y. Mechref, "Glycomic Profiling of Tissue Sections by LC-MS," *Anal. Chem.*, vol. 85, no. 8, pp. 4074–4079, 2013.

[117] Y. J. Chen, D. R. Wing, G. R. Guile, R. a Dwek, D. J. Harvey, and S. Zamze, "Neutral N-glycans in adult rat brain tissue--complete characterisation reveals fucosylated hybrid and complex structures.," *Eur. J. Biochem.*, vol. 251, no. 3, pp. 691–703, Feb. 1998.

[118] T. W. Powers, E. E. Jones, L. R. Betesh, P. R. Romano, P. Gao, J. A. Copland, A. S. Mehta, and R. R. Drake, "Matrix Assisted Laser Desorption Ionization Imaging Mass Spectrometry Workflow for Spatial Profiling Analysis of N-Linked Glycan

Expression in Tissues," *Anal. Chem.*, vol. 85, no. 20, pp. 9799–9806, 2013.

[119] S. Toghi Eshghi, X. Li, and H. Zhang, "Targeted analyte detection by standard addition improves detection limits in matrix-assisted laser desorption/ionization mass spectrometry.," *Anal. Chem.*, vol. 84, no. 18, pp. 7626–32, Sep. 2012.

[120] J. M. Spraggins and R. M. Caprioli, "High-speed MALDI-TOF imaging mass spectrometry: rapid ion image acquisition and considerations for next generation instrumentation.," *J. Am. Soc. Mass Spectrom.*, vol. 22, no. 6, pp. 1022–31, Jun. 2011.

[121] L. A. Klerk, A. F. M. Altelaar, M. Froesch, L. A. McDonnell, and R. M. A. Heeren, "Fast and automated large-area imaging MALDI mass spectrometry in microprobe and microscope mode," *Int. J. Mass Spectrom.*, vol. 285, no. 1–2, pp. 19–25, Aug. 2009.

[122] H.-R. Aerni, D. S. Cornett, and R. M. Caprioli, "Automated acoustic matrix deposition for MALDI sample preparation.," *Anal. Chem.*, vol. 78, no. 3, pp. 827–34, Feb. 2006.

[123] D. L. Baluya, T. J. Garrett, and R. A. Yost, "Automated MALDI matrix deposition method with inkjet printing for imaging mass spectrometry.," *Anal. Chem.*, vol. 79, no. 17, pp. 6862–7, Sep. 2007.

[124] J. Franck, K. Arafah, A. Barnes, M. Wisztorski, M. Salzet, and I. Fournier, "Improving tissue preparation for matrix-assisted laser desorption ionization mass spectrometry imaging. Part 1: using microspotting.," *Anal. Chem.*, vol. 81, no. 19, pp. 8193–202, Oct. 2009.

[125] L. A. McDonnell, A. van Remoortere, R. J. M. van Zeijl, H. Dalebout, M. R.

Bladergroen, and A. M. Deelder, “Automated imaging MS: Toward high throughput imaging mass spectrometry.,” *J. Proteomics*, vol. 73, no. 6, pp. 1279–82, Apr. 2010.

[126] J. Kiernan, *Histological and Histochemical Methods: Theory and Practice*. 1981.

[127] J. Chen, S. T. Eshghi, G. S. Bova, Q. K. Li, X. Li, and H. Zhang, “Epithelium percentage estimation facilitates epithelial quantitative protein measurement in tissue specimens Epithelium percentage estimation facilitates epithelial quantitative protein measurement in tissue specimens,” *Clin. Proteomics*, vol. 10:18, 2013.

[128] D. J. Harvey, “Proteomic analysis of glycosylation: structural determination of N- and O-linked glycans by mass spectrometry.,” *Expert Rev. Proteomics*, vol. 2, no. 1, pp. 87–101, Jan. 2005.

[129] B. L. Parker, M. Thaysen-Andersen, N. Solis, N. E. Scott, M. R. Larsen, M. E. Graham, N. H. Packer, and S. J. Cordwell, “Site-specific glycan-peptide analysis for determination of N-glycoproteome heterogeneity,” *J. Proteome Res.*, vol. 12, no. 12, pp. 5791–5800, 2013.

[130] S. Hua, C. C. Nwosu, J. S. Strum, R. R. Seipert, H. J. An, A. M. Zivkovic, J. B. German, and C. B. Lebrilla, “Site-specific protein glycosylation analysis with glycan isomer differentiation,” *Anal. Bioanal. Chem.*, vol. 403, no. 5, pp. 1291–1302, 2012.

[131] J. S. Strum, C. C. Nwosu, S. Hua, S. R. Kronewitter, R. R. Seipert, R. J. Bachelor, H. J. An, and C. B. Lebrilla, “Automated assignments of N- and O-site specific glycosylation with extensive glycan heterogeneity of glycoprotein mixtures,” *Anal.*

Chem., vol. 85, no. 12, pp. 5666–5675, 2013.

[132] J. Elias and S. Gygi, "Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry," *Nat. Methods*, vol. 4, no. 3, pp. 207–214, 2007.

[133] M. P. Campbell, R. Peterson, J. Mariethoz, E. Gasteiger, Y. Akune, K. F. Aoki-Kinoshita, F. Lisacek, and N. H. Packer, "UniCarbKB: building a knowledge platform for glycoproteomics.," *Nucleic Acids Res.*, vol. 42, no. Database, pp. D215–21, Jan. 2014.

[134] C. Cooper, E. Gasteiger, and N. Packer, "GlycoMod–a software tool for determining glycosylation compositions from mass spectrometric data," *Proteomics*, vol. 1, no. 2, pp. 340–349, 2001.

[135] K. Julenius, "Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites," *Glycobiology*, vol. 15, no. 2, pp. 153–164, 2004.

[136] K. Vosseller, "O-Linked N-Acetylglucosamine Proteomics of Postsynaptic Density Preparations Using Lectin Weak Affinity Chromatography and Mass Spectrometry," *Mol. Cell. Proteomics*, vol. 5, no. 5, pp. 923–934, 2006.

[137] Z. Darula and K. F. Medzihradszky, "Affinity enrichment and characterization of mucin core-1 type glycopeptides from bovine serum.," *Mol. Cell. Proteomics*, vol. 8, no. 11, pp. 2515–2526, 2009.

[138] Z. Darula, J. Sherman, and K. F. Medzihradszky, "How to Dig Deeper? Improved Enrichment Methods for Mucin Core-1 Type Glycopeptides," *Mol. Cell. Proteomics*, vol. 11, no. 7, pp. 1–10, 2012.

[139] C. Steentoft, S. Y. Vakhrushev, M. B. Vester-Christensen, K. T.-B. G. Schjoldager, Y. Kong, E. P. Bennett, U. Mandel, H. Wandall, S. B. Levery, and H. Clausen, "Mining the O-glycoproteome using zinc-finger nuclease–glycoengineered SimpleCell lines," *Nat. Methods*, vol. 8, no. 11, pp. 977–982, 2011.

[140] C. D. Wenger and J. J. Coon, "A Proteomics Search Algorithm Specifically Designed for High- Resolution Tandem Mass Spectra," *J. Proteome Res.*, vol. 12, pp. 1377–1386, 2013.

[141] A. Halim, U. Westerlind, C. Pett, M. Schorlemer, U. Rüetschi, G. Brinkmalm, C. Sihlbom, J. Lengqvist, G. Larson, and J. Nilsson, "Assignment of Saccharide Identities through analysis of oxonium ion fragmentation profiles in LC-MS/MS of glycopeptides.," *J. Proteome Res.*, vol. 13, no. 12, pp. 6024–32, 2014.

[142] M. G. Yet, C. C. Q. Chin, and F. Wold, "The covalent structure of individual N-linked glycopeptides from ovomucoid and asialofetuin," *J. Biol. Chem.*, vol. 263, no. 1, pp. 111–122, 1988.

[143] P. Hägglund, J. Bunkenborg, F. Elortza, O. N. Jensen, and P. Roepstorff, "A new strategy for identification of N-glycosylated proteins and unambiguous assignment of their glycosylation sites using HILIC enrichment and partial deglycosylation," *J. Proteome Res.*, vol. 3, no. 3, pp. 556–566, 2004.

[144] L. J. Huang, J. H. Lin, J. H. Tsai, Y. Y. Chu, Y. W. Chen, S. L. Chen, and S. H. Chen, "Identification of protein O-glycosylation site and corresponding glycans using liquid chromatography-tandem mass spectrometry via mapping accurate mass and retention time shift," *J. Chromatogr. A*, vol. 1371, pp. 136–145, 2014.

[145] J. E. Hansen, O. Lund, N. Tolstrup, A. A. Gooley, K. L. Williams, and S. Brunak,

"NetOglyc: Prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility," *Glycoconj. J.*, vol. 15, no. 2, pp. 115–130, 1998.

[146] D. L. Swaney, G. C. McAlister, and J. J. Coon, "Decision tree-driven tandem mass spectrometry for shotgun proteomics.," *Nat. Methods*, vol. 5, no. 11, pp. 959–64, 2008.

[147] S. R. Ellis, A. L. Bruinen, and R. M. A. Heeren, "A critical evaluation of the current state-of-the-art in quantitative imaging mass spectrometry," *Anal. Bioanal. Chem.*, vol. 406, no. 5, pp. 1275–1289, 2014.

[148] A. Varki, "Sialic acids in human health and disease.," *Trends Mol. Med.*, vol. 14, no. 8, pp. 351–60, Aug. 2008.

[149] M. Hedlund, E. Ng, A. Varki, and N. M. Varki, "alpha 2-6-Linked sialic acids on N-glycans modulate carcinoma differentiation in vivo.," *Cancer Res.*, vol. 68, no. 2, pp. 388–94, Jan. 2008.

[150] M. Asada, K. Furukawa, K. Segawa, T. Endo, and A. Kobata, "Increased Expression of Highly Branched N -Glycans at Cell Surface Is Correlated with the Malignant Phenotypes of Mouse Tumor Cells," *Cancer Res.*, vol. 57, pp. 1073–1080, 1997.

[151] M. Bucknall, K. Y. C. Fung, and M. W. Duncan, "Practical Quantitative Biomedical Applications of MALDI-TOF Mass Spectrometry," *Am. Soc. Mass Spectrom.*, vol. 13, no. 9, pp. 1015–1027, 2002.

[152] D. A. Pirman and R. A. Yost, "Quantitative Tandem Mass Spectrometric Imaging of Endogenous Acetyl-L-carnitine from Piglet Brain Tissue Using an Internal Standard," *Am. Soc. Mass Spectrom.*, vol. 21, no. 4, pp. 564–571, 2010.

[153] R. F. Reich, K. Cudzilo, J. A. Levisky, and R. A. Yost, "Quantitative MALDI-MS(n) analysis of cocaine in the autopsied brain of a human cocaine user employing a wide isolation window and internal standards.," *J. Am. Soc. Mass Spectrom.*, vol. 21, no. 4, pp. 564–71, Apr. 2010.

[154] L. Sleno and D. Volmer, "Assessing the properties of internal standards for quantitative matrix-assisted laser desorption/ionization mass spectrometry of small molecules," *Rapid Commun. mass Spectrom.*, vol. 20, pp. 1517–1524, 2006.

Curriculum Vitae

Shadi Toghi Eshghi was born in Isfahan, Iran, on September 21, 1987. Shadi graduated with honors in 2010 from Isfahan University of Technology, where she received her bachelor of science in Electrical and Computer Engineering. She enrolled at the Johns Hopkins School of Medicine in 2010 to pursue her Ph.D. in Biomedical Engineering. At Hopkins, Shadi joined the Center for Biomarker Discovery and Translation focusing her research on development of novel experimental and computational mass spectrometry techniques for isolation and identification of glycans and glycoproteins in biological and clinical samples. Her research has led to over 10 patents and journal papers, 15 conference presentations and several scientific awards including the Siebel Scholarship and the Society for Glycobiology Young Scientist Travel Award. Beyond research, Shadi is passionate about promoting women in STEM fields and leadership roles. She co-founded GWEN at Hopkins to empower women graduates in STEM through organizing professional development workshops, leadership series, and social events and raising awareness through publication of newsletters highlighting efforts and programs making positive change. Starting in April 2016, Shadi will join the Biomarker Development Department at Genentech in San Francisco, CA.